# Mathematical modeling of population dynamics, applications to vector control of Aedes spp. (Diptera:Culicidae)

Martin Strugarek

Sorbonne Université
Ecole doctorale Sciences Mathématiques Paris-Centre
*Laboratoire Jacques-Louis Lions*

# Modélisation mathématique de dynamiques de populations, applications à la lutte anti-vectorielle contre *Aedes spp.* (Diptera:Culicidae)

par Martin Strugarek

**Thèse de doctorat de Mathématiques Appliquées**

Dirigée par Nicolas Vauchelet et Benoit Perthame

Présentée et soutenue publiquement le 7 septembre 2018

devant un jury composé de :

| | | |
|---|---|---|
| M. | Vincent Calvez ..... | Examinateur |
| Mme | Marie Doumic ..... | Examinateur |
| M. | François Hamel ..... | Rapporteur |
| Mme | Anna Marciniak-Czochra ..... | Rapporteur |
| M. | Benoit Perthame ..... | Directeur de Thèse |
| M. | Lionel Roques ..... | Examinateur |
| M. | Nicolas Vauchelet ..... | Directeur de Thèse |

**Résumé.** On modélise la dynamique temporelle de populations de moustiques soumises à des interventions humaines par des systèmes déterministes, possédant ou non une structure spatiale, compartimentale ou en phénotype. En particulier, l'étude se concentre sur deux types d'interventions reposant sur des lâchers de moustiques appartenant à la même espèce que la population sauvage : mâles incompatibles seuls, en vue de l'élimination de population, ou bien mâles et femelles ensemble en vue de la modification de population, les individus relâchés présentant alors un autre phénotype que la population sauvage. Ces méthodes visent d'une part la réduction de la nuisance causée par les moustiques là où elle est la plus forte, et surtout, d'autre part, la diminution voire l'arrêt de la circulation des maladies infectieuse dont l'agent pathogène est transmis par leurs piqûres.

Les résultats mathématiques portent : d'abord sur le comportement asymptotique des solutions de systèmes paraboliques modélisant la dynamique de la fréquence d'individus qui présentent le phénotype introduit, dans le cas de la modification de population; puis sur une propriété qualitative (convergence vers un cycle limite périodique) des solutions de systèmes d'équations différentielles ordinaires modélisant des populations structurées en compartiments; ensuite sur le contrôle optimal par des lâchers d'individus d'un système d'équations différentielles ordinaires modélisant la modification de population en milieu homogène ainsi que le contrôle vers 0 d'un système modélisant des lâchers de mâles incompatibles pour l'élimination de population; et enfin, sur l'évolution de la structure en phénotype d'une population sexuée.

Ces résultats sont spécifiés aussi souvent que possible à des paramétrisations issues de données expérimentales, et illustrés par des simulations numériques. Leur interprétation pratique et leur éventuelle importance pour l'application sont systématiquement mises en lumière.

**Mots-clefs:** Dynamique de populations; lutte anti-vectorielle; modélisation; analyse asymptotique; réaction-diffusion; *Wolbachia*; *Aedes*; contrôle

**Abstract.** Deterministic systems are used to model time dynamics of mosquito populations undergoing human intervention. The models can have a spatial, compartmental or phentoypical structure. This study focuses on two kinds of intervention relying on releases of mosquitoes from the same species as the wild population: incompatible males for population elimination or males-and-females together for population replacement (there the released individuals have a phenotype different from that of the wild population). These methods aim at reducing the nuisance in highly infested areas and more importantly at limiting (or even stopping) vector-borne disease circulation.

Mathematical result are concerned with: asymptotic behavior of solutions to parabolic systems modeling the frequency of the introduced phenotype in the population, motivated by population replacement; a qualitative property of solutions to some ordinary differential systems (convergence to a periodic limit cycle) stemming from a compartmental structure; optimal control by males-and-females releases of an ordinary differential system modeling population replacement in a homogeneous environment, and the control to 0 of a model of population elimination by incompatible males releases; lastly time dynamics of the phenotypical structure in a sexual population.

As often as possible these results are specified using experimental data parametrization and illustrated through numerical simulations. Practical conclusions are drawn and the relevance with respect to the application is systematically highlighted.

**Keywords:** Population dynamics; vector control; modeling; asymptotic analysis; reaction-diffusion; *Wolbachia*; *Aedes*; control

# Remerciements

Cette thèse n'aurait pas existé si je n'avais pas rencontré plusieurs personnes passionnées et sachant transmettre leur passion, qui m'ont orienté vers les mathématiques. Je nomme ici mes professeurs MM. Pilloy, Bozec, Dupont et Pommellet, ainsi que mes frères Cyrille et Antoine. C'est à une séance de TD d'Yvan Martel à l'Ecole polytechnique que je dois mon initiation à la dynamique des populations. Qu'ils soient tous vivement remerciés.

Merci tout particulièrement à Marie Doumic, pour m'avoir présenté ceux qui allaient devenir mes directeurs de thèse, pour son accueil et ses conseils, et aussi pour son parcours d'IPEF et chercheuse en mathématiques appliquées qui m'a inspiré. Merci beaucoup de me faire l'honneur de participer à mon jury de thèse.

Je tiens à remercier chaleureusement mes deux directeurs de thèse qui ont proposé ce sujet et accepté de m'accompagner dans une démarche singulière. Ils se sont toujours montrés disponibles et bienveillants. Leur confiance m'a permis d'explorer à ma guise des pistes de recherche parfois inattendues, et leurs conseils précieux m'ont été d'un grand secours. Merci à Benoit Perthame pour les séances de travail au tableau où j'ai pu m'imprégner un peu de sa vision des objets mathématiques, toujours éclairante. Et un grand merci à Nicolas Vauchelet. Son humeur constante, sa persévérance et son souci permanent d'expliciter les ponts entre différents résultats mathématiques ont été des sources de motivation et d'inspiration.

C'est un honneur pour moi que François Hamel et Anna Marciniak-Czochra aient accepté de rapporter ma thèse. Merci aux chercheurs qui ont accepté de faire partie du jury, Vincent Calvez et Lionel Roques.

Cette thèse a été ponctuée de séjours en France ou à l'étranger, et je veux ici exprimer ma gratitude envers celles et ceux qui m'ont accueilli, et qui sont souvent devenus des collaborateurs. Warmful thanks to Jorge P. Zubelli, Daniel A. M. Villela, Maria Soledad Aronna, Claudia T. Codeço and Bento Pereira in Rio de Janeiro. You introduced me to various aspects of research, mathematics, entomology and Brazil. Jorge, I shall not forget your kindness and hospitality. Merci à Hervé Bossin en Polynésie française et Yves Dumont à Montpellier et Pretoria.

Merci également aux doctorants qui m'ont permis de présenter mes travaux lors de séminaires à Versailles, au Havre et à Montpellier : Hugo, Alexandre et Quentin.

Mes sincères remerciements vont aussi à Vincent Robert et Anna-Bella Failloux, qui m'ont fait confiance en acceptant un mathématicien dans leur cours d'entomologie médicale à l'Institut Pasteur. Merci pour cette expérience inoubliable et votre passion communicative, et merci à tous les élèves de la promotion Jean Mouchet, pour leur accueil et leur enthousiasme.

Merci à tous les membres permanents du LJLL avec qui j'ai eu la chance d'interagir, notamment à Catherine Drouet pour le manuscrit et Malika Larcher pour les missions. Merci à Luis Almeida pour son écoute, sa disponibilité et ses conseils précieux. Merci à Grégoire Nadin et Yannick Privat pour le travail en commun et tout ce que vous m'avez appris ! Merci spécialement à Pierre-Alexandre Bliman pour nos discussions toujours instructives (aussi bien à Paris qu'à Rio, Asuncion ou Porquerolles), pour m'avoir introduit aux systèmes monotones, pour la conférence MMCCD et pour le projet Mosticaw.

Merci aux doctorants du bureau 16-26 324, Ethem pour la passation de témoin et bien sûr Hongjun, avec qui j'ai eu beaucoup de plaisir à échanger sur la France, la Chine et les nourrissons, entre autres. Merci à Ana et Idriss, pour la fraîcheur que vous avez su apporter au bureau : il vous appartient de continuer à faire vivre ce bon esprit ! Merci aux autres doctorants du laboratoire avec qui j'ai interagi, en particulier Cécile et Camille en maths-bio.

Merci aux auteurs dont les livres m'ont nourri et accompagné pendant la thèse, et dont certains sont cités.

Merci à mes amis, doctorants ou non, qui ont partagé certains aspects de cette thèse et ont

été un soutien constant : Marc, Alban, Ambroise, Keyvan, Jean-Michel. Je veux aussi rappeler ici la mémoire de Nicolas, mort brutalement à l'été 2017. J'aurais tant aimé que tu puisses voir l'aboutissement de ce travail qui t'avait intéressé dès le départ... Nous avions partagé nos émois esthétiques et mathématiques depuis la classe prépa. Au nom de notre amitié je te dédie ce manuscrit.

Merci à ma famille, à mes parents et grands-parents ("Ce qu'on fait avec amour réussit toujours"), à Clotilde et Guillaume qui ont toujours manifesté de l'intérêt pour ma recherche et ont su me soutenir et m'encourager.

Merci enfin à mon épouse Perrine, pour son amour et son soutien dans les hauts et les bas du travail de thèse, et à mon fils François, qui aura dû supporter un père parfois absent ou distrait : que tu puisses puiser dans ce travail ce qu'il faudra de ténacité et de passion pour aller à ton tour au bout de tes idées.

# Contents

**Résumé** **3**

**Abstract** **3**

**Remerciements** **5**

**1 Introduction en français** **13**
1.1 Aspects généraux . . . 13
1.2 Outils . . . 15
1.3 Présentation des principaux résultats . . . 16

**2 English introduction** **23**
2.1 General aspects . . . 23
2.2 Tools . . . 24
2.3 Presentation of the main results . . . 26

**I Context** **33**

**3 Mosquitoes and vector control** **35**
3.1 Bio-ecology and monitoring of *Aedes* mosquitoes . . . 35
3.2 Vector-borne diseases . . . 37
3.3 Two vector control methods by releases . . . 39

**4 Population dynamics modeling** **41**
4.1 Notations, framework and motivation . . . 41
4.2 Differential equations . . . 44
4.3 Reaction-diffusion equations . . . 46
4.4 Important auxiliary results . . . 53

**II Reaction-diffusion models** **57**

**5 Reduction to a single equation for some 2-by-2 systems** **59**
5.1 Introduction . . . 59
5.2 Setting of the problem for typical Lotka-Volterra systems . . . 60
5.3 Application to a biological example . . . 63
5.4 Proof of convergence . . . 65
5.5 Generalization of the result . . . 73
5.6 Conclusion and perspectives . . . 75

**6 Hindrances to bistable propagation: wave-blocking, wave-delaying** **77**
6.1 Introduction . . . 77
6.2 Main results . . . 79
6.3 A brief reminder on bistable reaction-diffusion in $\mathbb{R}$ . . . 83
6.4 Proofs for the infection-dependent population gradient model . . . 84
6.5 Proofs for the heterogeneous case: blocking waves and barrier sets . . . 88
6.6 Discussion and extensions . . . 102

**Appendices** **107**
6.A An additional result . . . 107

**7 Uncertainty quantification for the invasion success** **109**
7.1 Introduction . . . 109
7.2 Setting the problem: How to use a threshold property to design a release protocol? 111
7.3 Critical bubbles of non-extinction in dimension 1 . . . 116
7.4 Specific study of a relevant set of release profiles . . . 120
7.5 Numerical results . . . 128
7.6 Conclusion and perspectives . . . 128

**Appendices** **131**
7.A Uniqueness of the minimal radius . . . 131

**III Temporal models** **135**

**8 Oscillatory regimes in a simplified model of hatching enhancement by larvae** **137**
8.1 Models and their reduction . . . 138
8.2 Study of the reduced model . . . 140
8.3 The slow-fast oscillatory regime . . . 145
8.4 Hopf bifurcation . . . 150
8.5 Conclusion . . . 154

**Appendices** **155**
8.A Numerical tests . . . 155
8.B Slow-fast computations . . . 157
8.C Numerics close to bifurcation . . . 158

**9 Using sterilizing males to reduce or eliminate *Aedes* populations: insights from a mathematical model** **161**
9.1 Modeling and biological parameter estimation . . . 162
9.2 Theoretical study of the simplified model . . . 165
9.3 Numerical study . . . 172
9.4 Conclusion . . . 178

**Appendices** **179**
9.A Study of the steady states . . . 179
9.B Basin entrance time approximation . . . 183

**10 Optimal releases for population replacement strategies, application to *Wolbachia*** **189**
10.1 Introduction . . . 189
10.2 Toward an optimal control problem . . . 192
10.3 Analysis of Problem ($\mathcal{P}_{\text{full}}$) and numerics . . . 196
10.4 Conclusion . . . 199

**Appendices** **203**
10.A Proofs . . . 203
10.B Qualitative properties of the minimizers . . . 209

**11 Sharp seasonal threshold property for cooperative population dynamics with concave nonlinearities** **215**
11.1 Introduction . . . 215
11.2 Context and motivation . . . 216
11.3 Results . . . 217
11.4 Proofs . . . 219
11.5 Discussion and extensions . . . 225

**Appendices** **229**

11.A Proof of Theorem 11.1 . . . . . 229

**IV Other aspects** **231**

**12 Selection-mutation dynamics with sexual reproduction** **233**
12.1 Introduction . . . . . 233
12.2 Main results . . . . . 235
12.3 The model without mutations . . . . . 238
12.4 $BV$ estimates on the total population . . . . . 241
12.5 Lyapunov approach . . . . . 246
12.6 The Hamilton-Jacobi equation . . . . . 247
12.7 Conclusion and perspectives . . . . . 253

**13 Mathematical perspectives** **255**
13.1 Bubbles for elliptic systems . . . . . 255
13.2 Wave-delaying . . . . . 257
13.3 Time-scales and limits for controlled slow-fast dynamics . . . . . 260
13.4 Stationary distributions for sexual reproduction kernels . . . . . 262

**Conclusion** **265**

# List of Figures


4.1 Schematic phase diagram . . . 51

5.1 Convergence to the scalar solution. . . . 66
5.2 Bistable reaction terms. . . . 75
5.3 Convergence to the scalar solution with imperfect maternal transmission. . . . 76

6.1 Sign of the traveling wave speed. . . . 83
6.2 Blocking and passing fronts with two different constant values for logarithmic population gradients. . . . 84
6.3 Blocking or passing fronts with two different initial data. . . . 84
6.4 Blocking or passing fronts with two different non-constant profiles of logarithmic population gradients. . . . 85
6.5 Minimal blocking gradient. . . . 85
6.6 Critical population ratio between the edges of the heterogeneous area (left) and its limit as a jump transition (right). . . . 86
6.7 Sketch for a phase-plane argument. . . . 92
6.8 Values of the critical blocked fronts at the edges of the heterogeneous area. . . . 105
6.9 Blocking or passing fronts with two different constant values of logarithmic population gradient in the case of a two-populations system. . . . 106

7.1 Profile of the reaction terms. . . . 112
7.2 Two-dimensional time dynamics with three different sizes for the release area. . . . 117
7.3 Minimal invasion radii from energy and critical bubbles. . . . 119
7.4 The sum of two Gaussian profiles. . . . 124
7.5 Under-estimation of introduction success probability. . . . 129
7.6 Degraded under-estimation when losing the constant $2\sqrt{2\log(2)}$. . . . 130

8.1 Slow-fast phase diagrams. . . . 146
8.2 Density-dependent hatching rate. . . . 150
8.3 Supercritical Hopf bifurcation diagram . . . 151
8.A.1 Sample egg dynamics . . . 155
8.A.2 Sample 3-dimensional dynamics. . . . 156
8.C.1 Sample eggs and larvae dynamics (1). . . . 158
8.C.2 Sample eggs and larvae dynamics (2). . . . 159
8.C.3 Sample eggs and larvae dynamics (3). . . . 159
8.C.4 Larvae dynamics period $T_0$ in days (*left*) and larvae dynamics amplitude (Amp) in percentage of $\overline{L}$ (*right*), for different couples $(a, b)$. . . . 160

9.1 Two viewpoints on the separatrix. . . . 174
9.2 Elimination trajectory and a zoom in the last 30 days of treatment. . . . 177
9.3 Time dynamics of different ratios . . . 177

10.1 Phase portrait of system (10.1). . . . 193
10.2 Bistable profile. . . . 195
10.3 Rectangular shapes of optimal solutions. . . . 198
10.4 Sample numerical discrete minimizers. . . . 199
10.5 Numerical error assessment between the discrete solution of the full problem and of the limit one for $C = 0.75$. . . . 200

10.6 Numerical error assessment between the discrete solution of the full problem and of the limit one for $C = 0.15$. . . . . 201

# List of Tables

8.C.1 Steady states, period and amplitude of oscillations for $a = .1$ . . . . . 159
8.C.2 Steady states, period and amplitude of oscillations for $a = .25$ . . . . . 160
8.C.3 Steady states, period and amplitude of oscillations for $a = .5$ . . . . . 160

9.1 Parameter values for some populations of *Aedes polynesiensis* in French Polynesia at a temperature of $27°C$. . . . . 164
9.2 Conversion of biological parameters into mathematical parameters. . . . . 165
9.3 Numerical values fixed for the simulations. . . . . 172
9.4 CPU times for the numerical simulations . . . . . 173
9.5 Effort ratio for various parameter values. . . . . 173
9.6 Estimations of the entrance time into the basin of **0** using analytic formulae. . . . 174
9.7 Entrance time into the basin of **0** for constant release flux. . . . . 174
9.8 Total effort ratio to get into the basin of **0** for constant release flux. . . . . 175
9.9 Total effort ratio to get into the basin of **0** for impulsive releases. . . . . 176
9.10 Entrance time into the basin of **0** for impulsive releases. . . . . 176
9.11 Entrance time into the basin of **0** for weekly releases. . . . . 176
9.12 Final total female ratio when the trajectory enters the basin of **0**. . . . . 176

10.1 Parameters for the numerical resolution of $(\mathcal{P}_{\text{full}})$ . . . . . 198

# Chapter 1

# Introduction en français

> Quand même les différences très sensibles, que j'ai remarquées dans les diverses contrées où j'ai abordé, ne m'auraient pas empêché de me livrer à cet esprit de système, si commun aujourd'hui, et cependant si peu compatible avec la vraie philosophie, comment aurais-je pu espérer que ma chimère, quelque vraisemblance que je susse lui donner, pût jamais faire fortune ?
>
> Louis-Antoine de Bougainville, *Voyage autour du monde par la frégate la Boudeuse et la flûte l'Etoile.*

## 1.1 Aspects généraux

### 1.1.1 Organisation du mémoire

Cette thèse est divisée en quatre parties. La partie I est dédiée à la présentation du contexte, c'est-à-dire des travaux, méthodes et problèmes dans le cadre desquels le travail s'est inscrit. La partie II rassemble les travaux concernant les équations de réaction-diffusion, qui sont les seuls à avoir inclus une ou plusieurs dimensions spatiales tandis que la partie III regroupe les études de dynamiques temporelles en dimension finie, décrites par des équations différentielles ordinaires. Enfin, la partie IV contient d'une part une étude de dynamique de population structurée en phénotype (Chapitre 12) et d'autre part des perspectives mathématiques (Chapitre 13).

Dans le détail, les Chapitres 3 et 4 sont dédiés à la présentation du contexte applicatif (c'est-à-dire l'entomologie médicale) d'une part, et mathématique d'autre part. A partir du Chapitre 5 jusqu'au Chapitre 12, cette thèse expose des résultats nouveaux, dont certains ont déjà été publiés dans des revues scientifiques (les références sont précisées en début de chapitre le cas échéant et listées ci-dessous). Avant la conclusion, le Chapitre 13 est destiné à motiver ou orienter une éventuelle poursuite des travaux de thèse : elle expose plusieurs problèmes ouverts apparus durant ces travaux, dont certains sont partiellement résolus.

Trois articles ont été publiés et un soumis dans des revues scientifiques :

- le Chapitre 5 a été publié dans SIAM Journal on Applied Mathematics [211] ;
- le Chapitre 6 a été publié dans Journal of Mathematical Biology [176] ;
- le Chapitre 7 a été publié dans Mathematical Biosciences and Engineering [212] ;
- le Chapitre 8 a été soumis.

### 1.1.2 Mathématiques et entomologie

Le sujet de cette thèse de mathématiques appliquées est orienté vers l'interface avec l'entomologie, et plus particulièrement avec la lutte anti-vectorielle. C'est pourquoi il importe de décrire en premier lieu en quoi cette interface a consisté, et quels en ont été les fruits.

J'ai bénéficié (en tant que module de formation de mon école doctorale) du cours intensif de l'Institut Pasteur de Paris intitulé "Insectes vecteurs et transmission d'agents pathogènes", au

mois de mars 2016. Cette formation en entomologie médicale a été précieuse parce qu'elle a permis ensuite des échanges fructueux avec des chercheurs d'autres disciplines que la mienne. Par ailleurs, la première année et demie de thèse a été ponctuée de temps d'interaction forte avec des partenaires brésiliens de mes directeurs, à Paris et à Rio de Janeiro. Cette interaction s'est traduite par deux projets de recherche ayant abouti, qui constituent les Chapitres 7 et 8 de cette thèse. Les années 2016 et 2017 ont également vu se réaliser le projet STIC-AmSud Mosticaw, coordonné par Pierre-Alexandre Bliman (Inria) en France. Deux réunions, la première au Paraguay et la seconde en France, ont permis de créer des liens et un réseau avec des chercheurs sud-américains motivés par l'utilisation de *Wolbachia* chez les moustiques du genre *Aedes*. Par la suite, j'ai mis en place, grâce à Yves Dumont (rencontré dans le cadre du projet Mosticaw) et à la confiance de mes directeurs, une collaboration entre le Laboratoire Jacques-Louis Lions et l'unité d'entomologie de l'Institut Louis Malardé (Polynésie Française) qui, je l'espère, ira en se développant, et dont l'apport scientifique à cette thèse est constitué par le Chapitre 9.

Du point de vue des mathématiques appliquées, l'autre versant de l'interface à laquelle se situe cette thèse, j'ai bénéficié d'un environnement exceptionnel au sein du laboratoire Jacques-Louis Lions, qui s'est traduit par de multiples collaborations par lesquelles j'ai notamment pu acquérir ou affiner des connaissances techniques en analyse. Le Chapitre 5 est l'aboutissement d'un travail commencé durant mon stage de Master 2 (encadré par Benoit Perthame et Nicolas Vauchelet) et a servi de point d'entrée dans la thématique de la modélisation de la lutte anti-vectorielle. Les Chapitres 6, 10, 11 et 12 sont eux issus de collaborations diverses avec des membres du LJLL intéressés par les mathématiques appliquées à la biologie.

Compte-tenu de ce contexte, il s'est agi de construire et d'étudier des modèles mathématiques de dynamique de population pouvant s'appliquer à des problèmes pratiques soulevés par des techniques innovantes de lutte anti-vectorielle contre des moustiques appartenant au genre *Aedes*. Il faut préciser que certains travaux présentés ici ont été directement motivés par des entomologistes, d'autres ont été élaborés en dialogue constant avec eux, et d'autres enfin ont été menés avant tout pour leur intérêt mathématique, sans jamais perdre de vue leur motivation biologique.

Travailler à l'interface soulève bien des écueils, et d'abord celui déjà relevé par Bougainville dans son *Voyage autour du monde* [63, p. 19] :

> Je suis voyageur et marin, c'est-à-dire un menteur et un imbécile aux yeux de cette classe d'écrivains paresseux et superbes qui, dans l'ombre de leur cabinet, philosophent à perte de vue sur le monde et ses habitants, et soumettent impérieusement la nature à leurs imaginations. Procédé bien singulier, bien inconcevable de la part des gens qui, n'ayant rien observé par eux-mêmes, n'écrivent, ne dogmatisent que d'après des observations empruntées de ces mêmes voyageurs auxquels ils refusent la faculté de voir et de penser.

C'est ainsi qu'en faisant des mathématiques motivées par l'entomologie on s'expose à la critique justifiée des "hommes de terrain" dès lors qu'on est tenté de "soumettre impérieusement" le système biologique aux conclusions toutes théoriques de l'étude d'un modèle mathématique. Pour autant, on ne doit pas non plus quémander l'indulgence des mathématiciens pour des résultats incomplets ou des preuves hasardeuses au motif qu'elles seraient justifiés par une interprétation biologique. Entre ces deux écueils dangereux se situe la route qu'on a tenté de prendre ici, voie étroite et parfois éreintante dans laquelle on n'a de cesse de confronter le modèle à notre connaissance de la réalité, et de soumettre les résultats conjecturés à l'examen mathématique le plus rigoureux. Ce travail à l'interface, avec les questions nouvelles qu'il soulève et les points de vue originaux qu'il pousse à adopter, est aussi une richesse qui, je l'espère, pourra parfois transparaître à la lecture des pages qui suivent.

Dans le domaine de l'entomologie, la modélisation mathématique se trouve face à des difficultés bien spécifiques, ainsi que des opportunités considérables. Les données collectées sont bien minces pour espérer en tirer une description fine d'une population. Du fait de la petite taille des organismes concernés, il s'avère difficile de suivre l'évolution d'une population à l'échelle individuelle. Cependant, du fait de leurs capacités importantes de dissémination et leur capacité à atteindre et à se développer dans des zones difficiles d'accès, le périmètre géographique n'est que rarement bien délimité. Ces difficultés de terrain étant posées, on voit au Chapitre 3 que les moyens d'investigation à la disposition des entomologistes médicaux permettent néanmoins d'acquérir une assez bonne connaissance des populations, au moins relativement, et rendent déjà pertinentes les conclusions d'un modèle bien informé, à condition qu'elles restent prudentes.

Les quelques travaux rassemblés dans cette thèse ne se conçoivent que comme une étape, qu'on espère utile, dans la constitution de modèles mathématiques bien adaptés et bien compris permettant d'anticiper et d'optimiser les effets de techniques innovantes de lutte anti-vectorielle reposant sur des lâchers d'individus. On justifie au Chapitre 4 la recherche de modèle simples, voire simplistes, mettant l'accent sur quelques mécanismes clairement identifiés et tentant d'en définir précisément les effets combinés.

## 1.2 Outils

Avant de présenter les résultats obtenus, nous mettons en lumière quelques aspects importants des outils employés.

### 1.2.1 Modélisation

Ces travaux reposent sur des modèles mathématiques déterministes, approche qui est détaillée au Chapitre 4. La description de la population d'insectes à un instant donné se trouve ainsi réduite soit à un nombre fini de quantités pouvant être comprises comme des effectifs (équations différentielles ordinaires), soit à une (ou deux, dans le cas de deux sous-populations en interaction) densité spatiale (équations de réaction-diffusion), soit enfin à une densité phénotypique (équations différentielles ordinaires structurées).

Il est crucial, dans la constitution de ces différents modèles, de bien identifier à la fois son périmètre (les questions auxquelles on souhaiterait répondre) et les mécanismes qu'il doit prendre en compte. Notre choix a toujours été d'aller vers le modèle le plus simple intégrant ces deux facteurs (périmètre et mécanismes impliqués). Là où les questions posées le nécessitaient, pour comprendre la dispersion de l'infection par *Wolbachia* à grande échelle, on a ainsi utilisé des équations de réaction-diffusion (Partie II) prenant en compte l'espace. A l'inverse, lorsqu'on s'intéressait à une population locale, ou au moins homogène, on s'est contenté d'une description temporelle de la dynamique (Partie III).

### 1.2.2 Systèmes monotones

Sauf dans les Chapitres 8 et 12, les modèles de dynamiques de populations étudiés dans cette thèse peuvent être vus comme des systèmes dynamiques monotones. Cette propriété structurelle - préserver une relation d'ordre sur l'espace d'états au cours de l'évolution temporelle - a été théorisée et étudiée de façon parallèle par Hirsch et Matano dans les années 1980 (voir par exemple [114]). Une conséquence remarquable en est la convergence *générique* vers un équilibre (voir le Chapitre 4 pour une discussion détaillée).

Ainsi, les modèles considérés induisent des dynamiques stéréotypées (convergence vers un équilibre), qui peuvent être décrites en se contentant de connaître les états d'équilibre et leur stabilité locale. Par ailleurs, la structure de monotonie permet d'avoir recours à un outil mathématique très adapté et particulièrement simple : les sur- et sous-solutions.

Du point de vue de l'interprétation biologique, la monotonie est satisfaisante si les mécanismes considérés sont univoques. Dans le cas des moustiques qui sera décrit plus bas, on peut considérer qu'un œuf se contente d'éclore pour donner une larve, qui devient pupe (nymphe) puis émerge comme adulte (imago), et (s'il s'agit d'une femelle) que cet adulte pond de nouveaux œufs (voir le Chapitre 3). Tant que le cycle de vie peut être décrit en ces termes, il est naturel qu'un accroissement initial du nombre d'œufs, par exemple, induise à tout instant ultérieur un accroissement non seulement du nombre d'œufs, mais aussi du nombre de larves, de nymphes et d'adultes par rapport à ce qu'aurait été la population en l'absence de cet accroissement initial. Cependant, comme l'illustre le Chapitre 8, dans la nature les relations sont rarement univoques, et des interactions - ou des chaînes d'interactions - complexes, notamment non-linéaires, peuvent mettre à mal cette intuition de monotonie. Plus précisément, dans le modèle étudié au Chapitre 8, on fait l'hypothèse que les larves présentes dans le milieu ont tendance à augmenter le taux d'éclosion des œufs. Cette rétro-action très simple suffit - sous conditions sur les paramètres - à déstabiliser la population d'équilibre et à induire des oscillations stables de la taille de la population qui peuvent être aussi importantes qu'on veut. En particulier, le comportement stéréotypé de convergence vers l'équilibre est ici perdu au profit d'une convergence vers une solution périodique.

Un autre exemple de perte de monotonie est fourni par le Chapitre 12. Dans ce chapitre, on considère un modèle de population structurée en phénotype (motivé par l'étude du trait de résistance à un insecticide). Ce travail s'inspire notamment de l'article de Pierre-Emmanuel Jabin et Gaël Raoul [123] dans lequel la dynamique de sélection selon un trait phénotypique d'une population en interaction compétitive globale est décrite. Bien qu'il ne s'agisse pas à proprement parler d'un système monotone, il faut noter que la structure particulière de cette équation confère une propriété de convergence générique vers une distribution d'équilibre (un ESD ou "distribution évolutivement stable" dans le vocabulaire de la dynamique adaptative), qui peut être décrite par une fonction de Lyapunov, fournie dans [123] par la méthode de l'entropie relative. Le travail du Chapitre 12 reprend cette étude en la modifiant pour modéliser la reproduction sexuée. Dès lors, les sous-populations présentant deux traits phénotypiques distincts sont à la fois en interaction compétitive (pour l'accès aux ressources) et coopérative par la reproduction (pour la transmission de leur trait par hérédité) : il n'y a pas de structure de monotonie.

### 1.2.3 Bistabilité

Une caractéristique commune des deux situations de lutte anti-vectorielle modélisées dans le cadre de cette thèse (plus précisément aux Chapitres 7, 9 et 10) est la bistabilité[1]. Dans le remplacement de population par *Wolbachia* (voir Section 3.3.2) comme dans l'élimination de population par la technique de l'insecte stérile ou incompatible (TIS/TII, voir Section 3.3.1), la situation initiale est une population de moustiques établie de façon stable dans un écosystème. Il s'agit du premier des deux équilibres stables du système : l'équilibre "sauvage" ou "naturel". L'intervention humaine, par des lâchers répétés de mâles et femelles (dans le premier cas) ou seulement de mâles (dans le second cas) vise à atteindre un autre équilibre stable. Dans le cas du remplacement, il s'agit de la situation où tous les individus sont porteurs de *Wolbachia*. Dans le cas de l'élimination, il s'agit de l'extinction de la population, laquelle est supposée stable (au sens où l'immigration d'individus, si elle est assez faible, ne suffit pas à implanter une population).

Le cadre mathématique adopté permet une analogie complète entre les deux situations, au moins au niveau des dynamiques temporelles[2]. La bistabilité, dans ce contexte où depuis l'un des deux états stables on cherche, par une intervention humaine ciblée, à atteindre l'autre état stable, pose ainsi une question simple et naturelle : comment faire passer le système d'un équilibre à l'autre ? Autrement formulée : par où passe la frontière entre les deux bassins d'attraction, et comment l'atteindre ? Pour formuler cette question en termes mathématiques, il importe de modéliser convenablement l'action humaine, et donc de définir soigneusement le problème de contrôle associé.

### 1.2.4 Théorie du contrôle

L'outil puissant constitué par la théorie du contrôle (optimal) des systèmes dynamiques différentiels s'est alors imposé (comme l'illustre le Chapitre 10), pour introduire rigoureusement dans les modèles les termes décrivant l'action humaine sur la population de moustiques.

Ce formalisme simple, où l'on voit l'état de la population à un instant donné comme la variable sur laquelle on agit, permet de formuler mathématiquement un objectif concret de lutte anti-vectorielle. On obtient ainsi un problème d'optimisation dont l'inconnue est le contrôle, c'est-à-dire l'action humaine sur le système.

## 1.3 Présentation des principaux résultats

### 1.3.1 Systèmes de réaction-diffusion

Dans trois articles, publiés à deux ou à trois avec Nicolas Vauchelet (directeur de thèse), Grégoire Nadin et Jorge P. Zubelli ([211], [176] et [212]), l'étude s'est focalisée sur des systèmes de deux équations de réaction-diffusion particuliers, posés sur un domaine spatial infini $\Omega = \mathbb{R}^d$. En y

[1] On qualifie de **bistable** un système dynamique possédant exactement deux états d'équilibre stables.

[2] Cependant, même en dimension 2, il faut noter que la séparatrice entre les deux bassins d'attractions n'est pas tout à fait de la même nature topologique. Comme me l'a fait remarquer Jean-Pierre Françoise, il contient deux équilibres distincts et donc une orbite hétérocline dans le cas du remplacement de population, tandis qu'il ne contient qu'un seul équilibre dans le cas de l'élimination de population, voir les Chapitres 9 et 10 pour plus de détails.

adjoignant un contrôle sous la forme d'une mesure positive $u$, ces systèmes s'écrivent

$$\begin{cases} \partial_t n_1 - \nabla \cdot \big(A(x)\nabla n_1\big) = b_1 n_1 (1 - s_h \dfrac{n_2}{n_1+n_2})(1 - \dfrac{n_1+n_2}{K}) - d_1 n_1 \text{ dans } [0,T]\times\mathbb{R}^d, \\ \partial_t n_2 - \nabla \cdot \big(A(x)\nabla n_2\big) = b_2 n_2 (1 - \dfrac{n_1+n_2}{K}) - d_2 n_2 + u \text{ dans } [0,T]\times\mathbb{R}^d, \\ n_i(0,\cdot) = n_i^0 \geq 0 \, (i \in \{1,2\}). \end{cases} \tag{1.1}$$

Ici, $b_i$ (resp. $d_i$) désigne le taux de fécondité (resp. de mortalité) nette de la population $i$ (sauvage pour $i = 1$, porteuse de *Wolbachia* pour $i = 2$). $K > 0$ est la capacité de charge de l'environnement et $s_h \in [0,1]$ est le taux d'incompatibilité cytoplasmique (parfaite lorsque $s_h = 1$). Dans [211] (qui fait l'objet du Chapitre 5 de cette thèse), avec $u \equiv 0$ et pour $b_i = b_i^0/\epsilon$ on montre que lorsque $\epsilon \to 0$, la proportion de la population 2, $p := n_2/(n_1+n_2)$ converge vers la solution de

$$\begin{cases} \partial_t p - \nabla \cdot \big(A(x)\nabla p\big) = d_2 \dfrac{p(1-p)(p-\theta)}{s_h p^2 - (s_f + s_h)p + 1}, \\ p(0,\cdot) = n_2^0/(n_1^0+n_2^0), \end{cases} \tag{1.2}$$

où

$$s_f := 1 - \frac{b_2^0}{b_1^0}, \quad \delta := \frac{d_2}{d_1}, \quad \theta := \frac{s_f + \delta - 1}{\delta s_h},$$

sous l'hypothèse $s_f < s_h$ (c'est-à-dire pour une incompatibilité cytoplasmique assez forte et un impact sur la fécondité assez faible).

Ce résultat (Théorème 5.1) peut en réalité être étendu à des termes de réaction un peu plus généraux, notamment pour prendre en compte une transmission verticale (= de la femelle à sa descendance) imparfaite ou encore des effets plus généraux de la fréquence d'infection $p$ sur les différents termes (voir les hypothèses 5.2 et 5.3 pour l'énoncé général). En substance, il s'agit d'une réduction de dimension par l'utilisation d'une réaction rapide[3]. Sa justification repose sur des estimations *a priori* dans des espaces de Sobolev appropriés permettant de prouver de la compacité par un lemme de Lions-Aubin, puis la convergence par unicité de la solution du problème limite.

L'intérêt d'une telle convergence est multiple. D'abord mathématiquement, l'équation (1.2), scalaire, dispose d'une formulation variationnelle qui permet de simplifier considérablement l'étude asymptotique. On peut également décrire en détail des sous-solutions ("bubbles") qui s'avèrent utiles (dans [212], mais aussi dans [176]) pour obtenir des conditions de non-extinction.

D'autre part sur le plan de la modélisation, notre motivation principale pour l'étude de (1.1) provient de l'article de Barton et Turelli [29], où un modèle en proportion du type de (1.2) est directement introduit pour modéliser la propagation spatiale d'un trait "variant", en présence de ce que les auteurs nomment "effets cytoplasmiques analogues à l'effet Allee" (en particulier, l'incompatibilité cytoplasmique causée par *Wolbachia* chez des espèces du genre *Aedes* entre dans cette catégorie). Il nous a semblé opportun de comprendre dans quelle mesure un modèle scalaire tel que (1.2), en proportion seulement, est capable de représenter simultanément l'évolution de deux sous-populations en interaction. Notre résultat de convergence permet de montrer rigoureusement que le modèle scalaire est proche du système (1.1), tout en dévoilant qu'en dehors de la limite $\epsilon \to 0$, la correspondance avec (1.2) est nécessairement imparfaite.

Enfin, l'équation non contrôlée (1.2), classique et bien comprise, possède pour $A = \mathbf{I}$ une propriété remarquable lorsque la donnée initiale est localisée : l'établissement de la population est quasiment[4] équivalente à sa propagation, cette dernière s'effectuant asymptotiquement (dans toute direction) selon un profil d'onde progressive (voir Section 4.3.2 *infra*). C'est d'ailleurs cette propagation à vitesse et profil constants qui a en premier lieu motivé l'utilisation de tels modèles pour *Wolbachia*, sur la base d'observations de terrain (voir les travaux de Turelli et Hoffmann, [222] et [223]).

Il apparaît que la correspondance (1.1)-(1.2) fournit un cadre idoine pour l'étude mathématique de mécanismes variés, selon le schéma suivant :

1. Modélisation dans (1.1) ;

[3] Lorsque $d = 0$ (cf. Chapitre 10), on obtient d'ailleurs un système lent-rapide classique en dimension 2 (voir Section 4.2.2), pour lequel cette réduction de dimension revient à une projection de la dynamique sur la variété lente.

[4] Au sens où l'établissement doit se faire dans un domaine suffisamment grand.

2. Passage à la limite vers (1.2) ;
3. Etude mathématique complète du problème limite ;
4. Comparaison à des résultats numériques sur (1.1).

C'est ce schéma qui est adopté dans le Chapitre 10, dans le cas particulier $d = 0$ (c'est-à-dire homogène en espace, ce qui peut aussi être vu comme la solution de (1.1) issue d'une donnée initiale constante). On y étudie le comportement des solutions lorsque $u = u^\epsilon$ résout un problème de contrôle optimal.

Dans [176] (qui fait l'objet du Chapitre 6), on se limite à la dimension 1 d'espace et on étudie le problème scalaire (1.2). L'étude est motivée par une remarque développée dans [29] au sujet de l'influence des variations spatiales de la taille de la population. Dans le modèle scalaire (de la forme (1.2)), Barton et Turelli notent que si le flux de gènes (*gene flow*) est modélisé par un terme d'advection faisant intervenir la taille de la population, c'est-à-dire si $p$ satisfait

$$\partial_t p - \partial_{xx} p - 2\frac{\partial_x N}{N}\partial_x p = f(p), \tag{1.3}$$

où $N : \mathbb{R} \to \mathbb{R}_+$ est donnée et $f$ est un terme de réaction bistable, alors pour $\frac{\partial_x N}{N}$ assez grand, la propagation de l'onde progressive est stoppée - le signe de sa vitesse de propagation peut même être modifié. On étudie de façon complète le cas où $\frac{\partial_x N}{N} = C\mathbb{1}_{[-L,L]}(x)$ pour $C, L > 0$, pour lequel on obtient le Théorème 6.2. En substance, à $L$ fixé si $C$ est assez grand, ou bien à $C$ fixé si $L$ est assez grand, alors on montre qu'il y a blocage de l'onde progressive. Sinon, l'onde est simplement retardée.

La Proposition 6.4 décrit le comportement asymptotique, lorsque $L$ tend vers 0, du produit $LC_*(L)$, où $C_*(L)$ est la valeur critique à partir de laquelle le blocage a lieu. L'interprétation de ce résultat est la suivante : le long d'une direction de propagation de l'invasion, si la taille de la population augmente de façon trop rapide, ou augmente de façon assez soutenue sur une distance assez longue, alors l'invasion est bloquée. A la limite, un saut brutal de la taille de la population suffit à stopper l'invasion si et seulement si son ampleur, mesurée par le rapport $N_R/N_L$ entre la taille de la population après le saut et celle avant le saut, excède

$$[N]_{\text{crit}} := \Big(1 - \frac{\int_0^1 f(p)dp}{\int_0^\theta f(p)dp}\Big)^{1/4}.$$

Cette limite est pertinente pour l'application à des population d'insectes, dont la densité peut varier brutalement en fonction de la végétation ou de la proximité de l'eau, par exemple. Ainsi, l'introduction de ce *gene flow* permet de modéliser ce type d'hétérogénéité spatiale.

Le Théorème 6.1 énonce quant à lui que si $N$ est en réalité une fonction de $p$, alors la nature de l'équation (1.3) n'est pas différente de celle sans advection : il s'agit d'une équation de réaction-diffusion bistable. Cependant, le signe de la vitesse de l'onde progressive peut être modifié.

Dans [212], qui fait l'objet du Chapitre 7, l'étude est centrée sur les données initiales. Le problème auquel on cherche à répondre consiste à déterminer, en présence d'incertitudes, comment relâcher à un instant donné des moustiques infectés afin d'assurer l'établissement de l'infection - et par conséquent sa diffusion, les deux concepts étant équivalents dans ce cas du point de vue pratique.

Le périmètre spatial dans lequel s'effectuent les lâchers étant borné, on montre dans un premier temps que la probabilité de succès tend vers 1 lorsque le nombre de points de lâcher tend vers l'infini (Proposition 7.2). Ensuite, on construit des profils de référence à support compact (sous-solutions en dimension 1 à l'aide de la résolution du problème stationnaire elliptique en domaine borné; profils d'énergie négative en dimension quelconque), nommés "bubbles" ou "propagules" tels que si le profil d'infection est, à un instant donné, partout supérieur à une propagule, alors l'infection se maintiendra pour tout temps, et envahira tout le domaine (Théorème 7.1).

On prouve ensuite que ces conditions suffisantes d'invasion (se trouver au-dessus d'une propagule) sont très difficiles à obtenir sur une donnée initiale issue d'un seul point de lâcher. Pour des points multiples de lâchers induisant une distribution de population infectée sous forme de somme de gaussiennes, on donne des résultats analytiques permettant de quantifier la probabilité d'invasion, ces résultats étant illustrés par des simulations numériques.

Une conséquence remarquable de cette étude est la mise en évidence d'une taille optimale pour le domaine de lâcher : même en présence d'incertitudes, un expérimentateur peut chercher à

définir la taille de la zone dans laquelle les moustiques seront lâchés afin d'optimiser la probabilité de succès. Si les lâchers sont trop concentrés, l'effet de la diffusion sera trop important et l'infection n'atteindra pas l' étendue spatiale critique pour se propager. A l'inverse, si les lâchers sont dispersés alors l'infection n'atteindra nulle part la fréquence critique pour se propager.

Du point de vue des systèmes monotones, on a affaire à une frontière entre bassins d'attraction qui passe aussi bien par des profils faibles en valeur mais étendus en espace que par des profils très localisés en espace avec des valeurs typiquement fortes.

Ce travail est à rapprocher de [93], où les auteurs ont étudié numériquement l'effet de deux observables macroscopiques de la donnée initiale, la fragmentation et l'abondance, sur la probabilité d'établissement de la population, pour une équation de réaction-diffusion bistable. Leur conclusion est que la fragmentation est plutôt délétère, avec un effet différencié selon le niveau d'abondance. On retrouve donc dans le travail du Chapitre 7 des questionnements similaires, puisque la taille du domaine maximisant la probabilité d'invasion varie selon l'abondance permise, et que cette taille joue typiquement sur la fragmentation de la donnée initiale : un périmètre plus étendu favorise la fragmentation.

### 1.3.2 Systèmes d'équations différentielles ordinaires

La plupart des autres travaux menés durant la thèse ont concerné des systèmes d'équations différentielles ordinaires.

D'abord, dans le Chapitre 8 on étudie un système très simple de dimension 2, issu d'un modèle à compartiments. Les deux dimensions restantes après simplification modélisent les œufs et les larves, et la particularité du modèle consiste à supposer que le taux d'éclosion $h \geq 0$ dépend de la densité larvaire (les autres paramètres étant des constantes positives):

$$\begin{cases} \dfrac{dE}{dt} = bL - (h(L) + d_E)E, \\ \dfrac{dL}{dt} = h(L)E - (cL + d_L)L. \end{cases} \tag{1.4}$$

Sur la base de cette hypothèse simple, on met en évidence le fait que si $h' > 0$ (c'est-à-dire si la rétro-action est positive), alors le système (1.4) peut effectivement être déstabilisé[5] et ses solutions peuvent présenter des oscillations périodiques. Ces oscillations apparaissent sous la forme d'une bifurcation de Hopf selon un paramètre décrivant l'intensité de la rétro-action, $h'(\overline{L})$ (où $(\overline{L}, \overline{E})$ est un équilibre positif stable pour des valeurs assez faibles de $h'(\overline{L})$, Théorème 8.2), ou bien à l'aide d'une dynamique lente-rapide lorsqu'on considère que le stock d'œufs $\overline{E}$ est grand et varie lentement (Théorème 8.1).

La mise en évidence de la possibilité - sur le plan de la modélisation - de telles oscillations simplement causées par une rétro-action au niveau de l'éclosion fournit une première pierre à une éventuelle explication de fluctuations observées dans la nature et qui ne semblent pas pouvoir être expliquées par les variations environnementales (voir [119], [141]). Cette étude devrait être poursuivie afin de mieux décrire, dans les modèles compartimentaux plus élaborés que (1.4), la dynamique d'éclosion des œufs dans le genre *Aedes* : il s'agit d'un phénomène non-linéaire certainement crucial pour bien comprendre ces populations.

Les deux chapitres suivants s'intéressent à des problèmes opérationnels posés par des méthodes de lutte anti-vectorielle, qui peuvent être vus comme des questions de contrôle de systèmes dynamiques différentiels.

L'utilisation de la technique dite de l'insecte incompatible (TII, voir Section 3.3.1) motive l'étude présentée au Chapitre 9. Il s'agit, par des lâchers de mâles stérilisants (qui rendent stériles les femelles avec lesquelles ils s'accouplent), de réduire voire d'éliminer une population d'insectes nuisibles, en l'occurrence des moustiques appartenant au genre *Aedes*. Le modèle proposé, volontairement simple, permet néanmoins de prendre en compte la dynamique des œufs (qui est considérée comme importante pour les espèces appartenant à ce genre) ainsi qu'un éventuel effet Allee[6]

[5] A l'inverse, si $h' \leq 0$ (1.4) devient monotone coopératif et possède 1 ou 0 équilibre positif. S'il possède un équilibre positif alors celui-ci est globalement asymptotiquement stable dans $(\mathbb{R}_+^*)^2$, et sinon $(0,0)$ est globalement asymptotiquement stable. Ce comportement justifie qu'on parle de "déstabilisation" dans le cas d'une rétroaction positive.

[6] Lorsque la densité de mâles est faible, on considère que certaines femelles ne sont pas fécondées.

quantifié par un paramètre $\beta > 0$ :

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_E E\big(1-e^{-\beta(M+M_i)}\big)\dfrac{M}{M+M_i} - \mu_F F, \\ \dfrac{dM_i}{dt} = u(t) - \mu_i M_i. \end{cases} \tag{1.5}$$

Ici, $E$ modélise tous les stades immatures, $M$ est la densité de mâles et $F$ la densité de femelles *fertiles*, c'est-à-dire ayant été fécondée par un mâle compatible. $M_i$ est la densité de mâles incompatibles et $u$ le flux de mâles relâchés.

L'effet Allee se traduit mathématiquement par la stabilité locale pour (1.5) de l'état trivial $\mathbf{0}$ où la population est éteinte, même lorsque $u \equiv 0$. Il s'agit d'une nouveauté dans les modèles de ce type, qui permet d'aborder les questions naturelles posées par cette méthode (amplitude et nombre des lâchers à effectuer typiquement) d'un point de vue analytique et géométrique : on se ramène de fait à la description d'une séparatrice qui est une sous-variété de co-dimension 1.

Après avoir défini, sur la base de la littérature entomologique, une gamme de valeurs pour les paramètres du modèles, on montre dans un premier temps que la dynamique temporelle est de nature bistable en général, l'équilibre positif stable étant noté $\mathbf{E}_+$. Quelques propriétés élémentaires mais utiles de la séparatrice entre les bassins d'attraction de $\mathbf{0}$ et $\mathbf{E}_+$ sont alors décrits dans la Proposition 9.2. En particulier, on établit qu'à partir d'un certain nombre (fini) d'œufs ou de femelles fertiles, la population se rétablit nécessairement vers son équilibre sauvage lorsqu'on cesse les lâchers. Le système dynamique induit par (1.5) sur $(E, M, F, M_i)$ étant par ailleurs monotone par rapport au cône $\mathcal{K}^o := \mathbb{R}^3_+ \times \mathbb{R}_-$, on montre que la population est conduite vers l'extinction lorsqu'elle est soumise à un lâcher continu de mâles stérilisants d'amplitude assez grande, $M_i(t) \equiv M_i > M_i^{\text{crit}}$. On établit alors des bornes analytiques sur le temps que prend l'extinction de la population à l'aide de sur- et sous-solutions (Proposition 9.3). On obtient des estimations de même nature lorsque les lâchers sont de type impulsionnel périodique d'amplitude $\Lambda$ et de période $\tau$ (Propositions 9.6 et 9.7). On montre pour cela que $M_i(t)$ converge en temps long vers la $\tau$-périodisation de $t \mapsto \frac{\Lambda e^{-\mu_i t}}{1-e^{-\mu_i \tau}}$. Ces estimations permettent d'estimer de façon analytique en fonction des paramètres un nombre de lâchers suffisant pour assurer l'extinction, à $\Lambda$ et $\tau$ fixés. L'étude numérique d'un cas particulier (une population isolée d'*Aedes polynesiensis*) permet d'illustrer ces résultats et de conforter leur intérêt pratique.

Complémentaire du chapitre précédent, le Chapitre 10 adopte le point de vue de la théorie du contrôle sur un problème similaire : celui de diriger un système d'équations différentielles ordinaires vers un équilibre stable, en partant d'un autre équilibre stable. Les systèmes étudiés aux Chapitres 9 et 10 présentent en outre le point commun d'être monotones et munis d'un contrôle monotone : des lâchers de mâles stérilisants dans le premier cas, et de moustiques mâles et femelles porteurs de *Wolbachia* dans le second. Le système étudié est donné par (1.1) homogène en espace, c'est-à-dire :

$$\begin{cases} \dfrac{dn_1}{dt} = b_1 n_1 (1 - s_h \dfrac{n_2}{n_1+n_2})(1-\dfrac{n_1+n_2}{K}) - d_1 n_1, \quad n_1(0) = K(1-\dfrac{d_1}{b_1}), \\ \dfrac{dn_2}{dt} = b_2 n_2 (1-\dfrac{n_1+n_2}{K}) - d_2 n_2 + u, \quad n_2(0) = 0. \end{cases} \tag{1.6}$$

Le système contrôlé (1.6) est complété par un critère décrivant la proximité de l'état du système au temps $T > 0$ par rapport au remplacement de population, c'est-à-dire à l'établissement de la population 2 au détriment de la population 1 :

$$J(u) := \frac{1}{2} n_1(T)^2 + \frac{1}{2}\big(K(1-\frac{d_2}{b_2}) - n_2(T)\big)_+^2. \tag{1.7}$$

Le problème de minimisation sous contrainte associé à (1.6)-(1.7), pour $u \in L^\infty(0,T)$ tel que $0 \le u \le M$ et $\int_0^T u(t)dt \le C$ pour $C, M > 0$ donnés est alors noté $(\mathcal{P}_{\text{full}})$.

On montre rigoureusement (Proposition 10.2) que lorsqu'on suppose que les paramètres de fécondité sont grands, $b_i = b_i^0/\epsilon$ ($i \in \{1,2\}$) et que l'on considère la limite $\epsilon \to 0$, alors $(\mathcal{P}_{\text{full}})$

converge vers un problème réduit ($\mathcal{P}_{\text{reduced}}$) défini par la minimisation, par rapport au même contrôle $u$, de la fonctionnelle

$$J^0(u) := (1 - p(T))^2,$$

où $p$ est solution de l'équation suivante :

$$\frac{dp}{dt} = p(1-p)\frac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} + \frac{u}{K}\frac{b_1^0(1-p)(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p}, \quad p(0) = 0. \tag{1.8}$$

L'équation (1.8) est intéressante car elle permet de décrire précisément le transfert d'un contrôle à la modélisation simple (lâchers d'individus porteurs de *Wolbachia*, $u(t)$ est donc le débit de moustiques relâchés dans (1.6)) vers un système de contrôle scalaire portant sur la proportion d'individus porteurs de *Wolbachia* seulement.

Le problème ($\mathcal{P}_{\text{reduced}}$) peut quant à lui être résolu analytiquement (Théorème 10.1) : pour tout $M > 0$ il existe un $C^*(M) > 0$ exprimable en fonction des paramètres ($b_i^0$, $d_i$, $s_h$ et $K$) tel que si $C > C^*(M)$ la solution de ($\mathcal{P}_{\text{reduced}}$) est unique et donnée par $u^* = M\mathbb{1}_{[0,C/M]}$, si $C < C^*(M)$ cette solution est également unique et donnée par $u^* = M\mathbb{1}_{[T-C/M,T]}$, et si $C = C^*(M)$ alors l'ensemble des minimiseurs est donné par $u^*_\lambda = M\mathbb{1}_{[\lambda,\lambda+C/M]}$ pour $\lambda \in [0, T - C/M]$.

La combinaison de ces deux résultats (convergence et résolution du problème limite) permet d'affirmer que dans le cas où le remplacement de population est possible, et si les fécondités sont grandes, on est proche de l'optimalité en lâchant toute la population au plus grand débit possible, le plus tôt possible. L'étude numérique du problème ($\mathcal{P}_{\text{full}}$) permet toutefois de montrer qu'en dehors de la limite $\epsilon \to 0$, les stratégies optimales de remplacement diffèrent, parfois sensiblement, de la stratégie limite.

La partie III se clôt par le Chapitre 11, qui est une contribution théorique à l'étude de dynamiques saisonnières. La question de départ consiste à se demander quelles sont les conséquences de la prise en compte de saisons différenciées dans les dynamiques de populations étudiées plus haut. En premier lieu, on s'attache à décrire aussi bien que possible le cas où deux saisons alternent, l'une favorable et l'autre défavorable. On prouve (Théorème 11.2) que sous hypothèse de concavité des non-linéarités, on peut assez aisément trouver des conditions sous lesquelles la durée relative des deux saisons est un paramètre discriminant de façon simple la dynamique : soit la population s'éteint (si la saison défavorable est régulièrement trop longue), soit elle tend vers un unique profil périodique (si la saison défavorable est assez brève). Ce résultat préliminaire demande à être mis en rapport avec les problèmes de contrôle associés aux techniques de lutte anti-vectorielle modélisées (notamment par l'étude des propriétés du cycle limite périodique), et son extension à d'autres types de non-linéarités doit être étudiée (par exemple, alternance saisonnière de dynamiques bistable et monostable d'extinction).

### 1.3.3 Equation d'évolution pour une population structurée

Enfin, le Chapitre 12 étudie une population structurée en phénotype. L'état de la population y est décrit par une mesure positive sur un espace $\mathcal{P}$ de phénotypes, typiquement $\mathcal{P} \subseteq \mathbb{R}^d$. Le modèle étudié est motivé par le problème actuel posé par la dynamique de résistance aux insecticides dans les populations de moustiques. Le modèle d'équation différentielles ordinaires étudié par Schechtman et Souza dans [200] a permis aux auteurs de décrire l'asymétrie entre l'apparition de la résistance lors de l'exposition d'une population à un insecticide, typiquement sur une échelle de temps courte, et sa disparition lorsque l'insecticide n'est plus utilisé, sur une échelle de temps longue. Ce résultat a été obtenu sur un modèle simple de résistance causée par deux mutations successives, la première engendrant une résistance correcte pour un coût (en fitness, c'est-à-dire en adaptation au milieu) élevé, et la seconde renforçant la résistance tout en réduisant drastiquement son coût. Notre objectif est de décrire ce type d'asymétrie (ou d'autres dynamiques temporelles intéressantes) pour une population à structure continue, en tenant compte de la nature sexuée de la reproduction. Ceci constitue une nouveauté importante par rapport aux travaux de modélisation existants, dont beaucoup s'intéressent à la résistance aux traitements dans des populations de bactéries ou de cellules où la reproduction de type clonal domine (voir par exemple [153]).

Le modèle proposé est de facture générique, et le traitement est surtout mathématique. On vise à décrire des outils permettant de traiter le terme de reproduction sexuée, par nature non-linéaire et non-local, quoique homogène, pour une équation d'évolution de la forme suivante :

$$\partial_t n(t,x) = \frac{1}{\rho(t)} \iint_{\mathcal{P}^2} K(x,y,z) n(t,y) n(t,z) dy dz - R(x,\rho(t)) n(t,x), \quad \rho(t) = \int_{\mathcal{P}} n(t,x) dx. \tag{1.9}$$

Ici, $K(x,y,z) \geq 0$ est la distribution pondérée de la progéniture issue de la fécondation d'une femelle de trait $y$ par un mâle de trait $z$ et $R(x,\rho)$ est la mortalité des individus de trait $x$ lorsque la population totale est égale à $\rho$. Un tel modèle n'est justifié que dans la mesure où on suppose que le rapport entre mâles et femelles est constant, aussi bien en temps qu'entre les différents traits $x \in \mathcal{P}$. On suppose typiquement que $R(x,\cdot)$ est une fonction croissante et non bornée qui modélise la saturation du milieu, ce qui permet de borner $\rho$ uniformément le long des orbites.

Dans un premier temps, on décrit les propriétés particulières des modèles dits "imitatifs" (ce terme provient de la théorie des jeux, voir [199]), où la progéniture hérite toujours du trait exact de l'un de ses parents. Pour ces modèles particuliers, une fonctionnelle de Lyapunov peut être introduite dans certains cas, qui permet de prouver la concentration de la population vers le trait le plus adapté (Théorème 12.5). Ensuite, on s'intéresse à la dynamique en temps long, dans la limite où l'échelle des mutations (c'est-à-dire l'écart à une dynamique imitative) devient petite. Cette approche donne lieu à la construction (classique) d'un objet limite sous la forme d'une équation de Hamilton-Jacobi avec contrainte, qui présente ici un aspect nouveau, dû au terme de reproduction sexuée (Théorème 12.4 et Section 12.6). En supposant par exemple que $\epsilon K_\epsilon(x,y,z) = B(y)\alpha(\frac{x-z}{\epsilon},y)$ où $\int \alpha(z',y)dz' \equiv 1$, c'est-à-dire que la fécondité du croisement entre un mâle $z$ et une femelle $y$ ne dépend que de $y$, alors on obtient à la limite $\epsilon \to 0$ :

$$\partial_t u(t,x) = \int B(y)q(t,y)\mathcal{L}[\alpha(\,\cdot\,,y)](\partial_x u(t,x))dy - R(x,\rho(t)), \quad \max_{x\in\mathbb{R}} u(t,x) \equiv 0,$$

où $\mathcal{L}[\alpha(\,\cdot\,,y)]$ est la transformée de Laplace de $\alpha(\,\cdot\,,y)$ :

$$\mathcal{L}[\alpha](p) := \int \alpha(z)e^{-p\,\cdot\,z}dz,$$

et $q(t,y) = \lim_{\epsilon\to 0} n_\epsilon(t,y)/\rho_\epsilon(t)$ est la distribution (normalisée) de la population. L'étude de cette équation limite, inachevée dans l'état actuel de ce travail, est rendue difficile par la présence de termes dépendant de $t$, $R(x,\rho(t))$ d'une part et $q(t,y)$ d'autre part, qui doivent être définis précisément par passage à la limite dans l'équation au niveau $\epsilon$. Une première étape dans cette étude est l'obtention de bornes uniformes en variation totale locale sur $\rho_\epsilon(t)$, ce qui est l'objet des Théorèmes 12.1, 12.2 et 12.3, pour trois situations particulières. L'extension de ces résultats et méthodes à des noyaux de reproduction $K$ plus généraux est également discutée.

Les questions de l'existence et de l'unicité des états d'équilibres pour (1.9) sont proposées comme problèmes ouverts dans la Section 13.4.

# Chapter 2

# English introduction

> Quand même les différences très sensibles, que j'ai remarquées dans les diverses contrées où j'ai abordé, ne m'auraient pas empêché de me livrer à cet esprit de système, si commun aujourd'hui, et cependant si peu compatible avec la vraie philosophie, comment aurais-je pu espérer que ma chimère, quelque vraisemblance que je susse lui donner, pût jamais faire fortune ?
>
> Louis-Antoine de Bougainville, *Voyage autour du monde par la frégate la Boudeuse et la flûte l'Etoile.*

## 2.1 General aspects

### 2.1.1 Dissertation outline

This manuscript is split into four parts. Part I is devoted to context presentation (*i.e.* existing works, methods and problems which frame the present work). Part II gathers studies on reaction-diffusion equations. These are the only ones which include one or several spatial dimensions. Part III groups together works concerned with time dynamics in finite dimension described by ordinary differential equations. Finally Part IV contains the study of the dynamics of a populations with a phenotype structure (Chapter 12) and mathematical perspectives (Chapter 13).

In details, Chapters 3 and 4 are devoted to the application context (medical entomology) and to the mathematical context, respectively. From Chapter 5 to Chapter 12 this thesis presents new results, some of which have already appeared in scientific journals (the corresponding references are given at the beginning of each chapter) as listed below. Before conclusion, Chapter 13 is an attempt to motivate or guide possible extensions of the present work. Several open problems are stated (some of them being partly solves), which appeared during the thesis completion.

Three articles were published and one submitted in scientific journals:

- Chapter 5 was published in SIAM Journal on Applied Mathematics [211] ;
- Chapter 6 was published in the Journal of Mathematical Biology [176] ;
- Chapter 7 was published in Mathematical Biosciences and Engineering [212] ;
- Chapter 8 has been submitted.

### 2.1.2 Mathematics ans entomology

This thesis in applied mathematics interfaces with entomology, more precisely with vector control. Therefore, in the first place we describe what this interface was, and what have been its benefits.

In march 2016 I took the intensive course from Institut Pasteur de Paris on vector insects and pathogens transmission (as a module from my doctoral school). This training in medical entomology has been precious to nurture exchanges with researchers in other fields than mine. Moreover strong interactions with Brazilian partners of my advisors punctuated the first year and a half of the thesis, in Paris and Rio de Janeiro. This interaction brought about two research

projects that were completed to yield Chapters 7 and 8. In 2016 and 2017 also, the STIC-AmSud Mosticaw project was developed under the supervision of Pierre-Alexandre Bliman (Inria) for the French part. Two meetings in Paraguay and France made it possible to exchange and build a network with researchers from South America motivated by the use of *Wolbachia* to control mosquitoes in genus *Aedes*. Afterwards, with the help from Yves Dumont (met during Mosticaw meetings) and trust from my advisors I put in place a collaboration between Laboratoire Jacques-Louis Lions (LJLL) and the entomology unit from Institut Louis Malardé (French Polynesia). So far, the scientific outcome of this joint effort is embodied by Chapter 9, which I hope will be pursued.

The other facet of the interface are applied mathematics. In this area LJLL is an outstanding environment where I could work with various researchers and learn a lot technically from them especially in analysis. Chapter 5 is the completion of a study I began during my Master 2 internship (under the supervision of Benoit Perthame and Nicolas Vauchelet). It was the starting point for vector control modeling. Chapters 6, 10, 11 and 12 are joint works with members of LJLL interested in bio-mathematics.

Considering this context, I built and studied mathematical models of population dynamics which can be applied to practical problems raised by innovative vector control techniques directed against mosquitoes in genus *Aedes*. Some of the works presented here were directly motivated by entomologists, some other works were joint efforts in tight connection with them and yet other works were performed out of mathematical interest, while keeping in mind their biological motivation.

Interfaces are rich in pitfalls. Bougainville in *Voyage autour du monde* [63, p. 19] already noted (loose translation)

> I am a traveler and a sailor, that is a liar and a fool in the eyes of those lazy and haughty writers in the shadows of their cabinets who philosophize endlessly about the world and its inhabitants, imperatively submitting nature to their thoughts. What a singular, utterly unbelievable manner to those people: they haven't observed anything themselves but write and get dogmatic about observations they borrow to the very same travelers to whom they deny the ability to see or think.

When doing mathematics motivated by entomology, one get exposed to the criticism of "hands-on people" as soon as one gives way to the temptation of "submitting imperatively" a biological system to purely theoretical conclusions drawn from a mathematical model. Yet, one mustn't expect the leniency of mathematicians for incomplete or slippery proofs on behalf of biological interpretations. Right between these two pitfalls goes the way we tried to travel here. Walking this narrow and sometimes exhausting path one never ceases to bring model face to face with evidence-based knowledge, submitting conjectures to the most thorough mathematical analysis. The work at interface is also rich in new questions and original viewpoints. I hope that this will show through these pages.

When dealing with entomology the mathematical modelers are faced with specific problems and opportunities. The available data are too coarse (or costly, see Chapter 3) to hope for a very fine description of a given population. Due to their small size, it is hopeless to try and follow individual mosquitoes. However these insects can disperse, getting hard to reach: their geographic area is seldom well-defined. Considering these field difficulties, medical entomologists can still gain a pretty good knowledge (at least in relative terms) of populations thanks to their investigation tools (see Chapter 3). Therefore, humble conclusions from a well-used model can already prove relevant.

The works gathered in this dissertation are nothing but a step (a useful one, hopefully) in designing well-adapted and well-understood mathematical models that allow for anticipating and optimizing the outcomes of innovative vector control methods, relying on releases of individuals. In Chapter 4 we justify our quest for simple (or even simplistic) models which stress a few clearly identified mechanisms and can predict precisely the results of their combination.

## 2.2 Tools

Before presenting our results we highlight some important aspects of the tools involved.

### 2.2.1 Modeling

The present work is based upon deterministic mathematical models. The approach is detailed in Chapter 4. The insect population at a given moment in time is thus reduced either to a finite number of figures, the headcounts (ordinary differential equations), to one or two (in case of two interacting sub-populations) spatial densities (reaction-diffusion equations), or to a density over a phenotype space (structured ordinary differential equations).

When shaping such a model, two features are critical: its range (what questions would it answer) and the mechanisms it takes into account. We always targeted the simplest model incorporating these two features. When necessary, to understand *Wolbachia* infection dispersal at large scale we used reaction-diffusion equations (Part II) that take space into account. On the contrary, for local or homogeneous populations we sticked to temporal dynamics (Part III).

### 2.2.2 Monotone systems

Except in Chapters 8 and 12, all population dynamics models in this dissertation can be seen as monotone dynamical systems. The monotonicity property (to preserve an order on the state space throughout time evolution) is structural and was studied simultaneously by Hirsch and Matano in the 1980s (see for instance [114]). A remarkable consequence is *generic* convergence to equilibrium (see Chapter 4 for a detailed account).

Thus, the models we consider induce standard dynamics (convergence to equilibrium), which require only the knowledge of the steady states and their local stability. In addition, monotonicity provides a very convenient and simple mathematical tool: sub- and super-solutions.

From the biological point of view, monotonicity is satisfactory if the mechanisms are unequivocal. Let us expand on the mosquito case. At first sight, an egg hatches and produces a larva, which becomes a pupa and emerges as an adult; if it is a female, it will lay new eggs (see Chapter 3). As long as the life-cycle goes along these lines, it makes sense that any increase in the initial number of eggs, for instance, results in an increase not only of the number of eggs but also of the numbers of larvae, pupae and adults compared to what they would have been without this increase. However, Chapter 8 illustrates the fact that natural interactions are almost never unequivocal. Complex, nonlinear interactions (or interaction chains) jeopardize the intuitive monotonicity. In the model from Chapter 8, we hypothesize that larvae are a cue for egg hatching. This very simple feedback is enough (under some parameter conditions) to destabilize the equilibrium population and induce stable oscillations of the population sizes, which can be arbitrarily large. In particular the "generic convergence to equilibrium" behavior disappears, it is replaced by the convergence to a periodic solution.

Monotonicity is also lost in Chapter 12. In this chapter, a population is structured by its phenotype. The study is motivated by the insecticide resistance trait and is inspired by the paper [123] by Pierre-Emmanuel Jabin and Gaël Raoul. The authors describe the selection dynamics of a trait under global competitive interactions in the population. Although this is not *stricto sensu* a monotone system, the singular structure of the equation yields the generic convergence to equilibrium property (in fact, toward an ESD for "evolutionarily stable distribution"), which can be proved by using a Lyapunov function relying on the relative entropy in [123]. In Chapter 12 we modify the model to take into account sexual reproduction. Any two sub-populations with different traits are then simultaneously in competitive interaction (for resources) and in cooperative interaction by reproduction (for the transmission of their trait by heredity): there is no monotonicity.

### 2.2.3 Bistability

The two vector control methods we model here (more precisely in Chapters 7, 9 and 10) share a bistable[1] nature. In *Wolbachia* population replacement (see Section 3.3.2) as in population elimination by sterile or incompatible insect technique (SIT/IIT, see Section 3.3.1), the initial state consists in an established wild mosquito population: this it the first stable steady state. Human interventions repeatedly release males and females (for replacement) or males only (for elimination). It aims at another stable steady state: established *Wolbachia* infection in the first case and population eradication in the second case (which is assumed to be stable in the sense that immigrants cannot implant a population if they are too few).

[1] A dynamical system is termed **bistable** when it has exactly two stable steady states.

Our mathematical framework stresses the analogy between the two situations, at least for temporal dynamics[2]. Since we use a human intervention to make the system go from a stable steady state to another stable steady state, bistability raises a simple and natural question: how to make the system move from a steady state to the other? In other words, where lies the boundary between the two basins of attraction, and how to reach it? To express this mathematically, the human intervention itself must be modeled properly: the associated control problem must be carefully defined.

### 2.2.4 Control theory

(Optimal) control theory for differential dynamical systems is a powerful tool, which was a must to tackle the previous problem. Chapter 10 illustrates the rigorous introduction of control terms in mosquito population dynamics models to describe human intervention.

In this simple formalism the population state at a given moment in time is the variable we act upon. A practical problem of vector control can then translate mathematically into an optimization problem whose unknown is precisely the human intervention on the system.

## 2.3 Presentation of the main results

### 2.3.1 Reaction-diffusion systems

In the three papers [211], [176] and [212] (co-authored with Nicolas Vauchelet, doctoral supervisor and Grégoire Nadin or Jorge P. Zubelli) we focused on special reaction-diffusion systems with two components posed on the infinite domain $\Omega = \mathbb{R}^d$. Adding a positive measure $u$ as a control term these systems read

$$\begin{cases} \partial_t n_1 - \nabla \cdot \big(A(x)\nabla n_1\big) = b_1 n_1 (1 - s_h \dfrac{n_2}{n_1+n_2})(1 - \dfrac{n_1+n_2}{K}) - d_1 n_1 \text{ in } [0,T]\times\mathbb{R}^d, \\ \partial_t n_2 - \nabla \cdot \big(A(x)\nabla n_2\big) = b_2 n_2 (1 - \dfrac{n_1+n_2}{K}) - d_2 n_2 + u \text{ in } [0,T]\times\mathbb{R}^d, \\ n_i(0,\,\cdot\,) = n_i^0 \geq 0 \, (i \in \{1,2\}). \end{cases} \tag{2.1}$$

Here, $b_i$ (resp. $d_i$) is the net fecundity (resp. death) rate of population $i$ (wild for $i = 1$, carrying *Wolbachia* for $i = 2$). $K > 0$ is the environmental carrying capacity, $s_h \in [0,1]$ is the cytoplasmic incompatibility (CI) rate (CI is perfect when $s_h = 1$). In [211] (which is Chapter 5), with $u \equiv 0$ and $b_i = b_i^0/\epsilon$ we show that the frequency of *Wolbachia* infection $p := n_2/(n_1+n_2)$ converges as $\epsilon \to 0$ to the solution of

$$\begin{cases} \partial_t p - \nabla \cdot \big(A(x)\nabla p\big) = d_2 \dfrac{p(1-p)(p-\theta)}{s_h p^2 - (s_f + s_h)p + 1}, \\ p(0,\,\cdot\,) = n_2^0/(n_1^0 + n_2^0), \end{cases} \tag{2.2}$$

where

$$s_f := 1 - \frac{b_2^0}{b_1^0}, \quad \delta := \frac{d_2}{d_1}, \quad \theta := \frac{s_f + \delta - 1}{\delta s_h},$$

under the assumption $s_f < s_h$ (*i.e.* if CI is strong enough and fecundity reduction is limited).

As a matter of fact this result (Theorem 5.1) extends to slightly more general reaction terms, including imperfect vertical (=from mother to offspring) transmission, or more general effects of the infection frequency $p$ on the various terms (the general statement holds under assumptions 5.2 and 5.3). In substance the dimension is reduced due to fast reaction[3]. It is proved by means of *a priori* estimates in suitable Sobolev spaces, yielding compactness thanks to a Lions-Aubin lemma and the full convergence by uniqueness of the solution to the limit problem.

[2] However, even in dimension 2 it must be emphasized that the separatrix between the two basins of attraction does not have the same topological nature. Jean-Pierre Françoise made me notice that in the case of population replacement the separatrix contains two distinct (unstable) steady states, hence a heteroclinic orbit, while it contains only one (unstable) steady state in the case of population elimination. See Chapters 9 and 10 for further details.

[3] When $d = 0$ (cf. Chapter 10), we obtain a classical slow-fast system in dimension 2 (see Section 4.2.2) for which this dimension reduction is equivalent to the projection of the dynamics on the slow manifold.

The interest of this convergence result is manifold. First mathematically: equation (2.2) is scalar and enjoys a variational formulation which simplifies a lot the asymptotic study. Some sub-solutions ("bubbles") can also be described in details and prove useful to get non-extinction conditions (in [212] and also in [176]).

Then from the modeling point of view, our starting motivation for studying (2.1) stems from the paper by Barton and Turelli [29] where a frequency model similar to (2.2) is used directly to model the dispersion of a variant trait in the presence of so-called "cytoplasmic and genetic analogues of Allee effects" (in particular, cytoplasmic incompatibility caused by *Wolbachia* is one of these effects). We found that it was worthy of interest to understand to which extend a scalar model such as (2.2), dealing only with infection frequency, is able to represent simultaneously the dynamics of two interacting sub-populations. Our convergence results shows rigorously that the scalar model is close to system (2.1) but also unveils the imperfect correspondence with (2.2) when $\epsilon > 0$.

Lastly the (uncontrolled) equation (2.2) is classical and well-understood. When $A = \mathbf{I}$ it possesses a remarkable property for localized initial data: establishment of the population is practically[4] equivalent to propagation, and in this case it propagates (along any direction) as a traveling wave asymptotically (see section 4.3.2 *infra*). It is precisely this constant-speed, constant-profile propagation of *Wolbachia* which aroused the interest for such models, based upon field observations (see the works of Turelli and Hoffmann [222] and [223]).

It appears that the correspondence (2.1)-(2.2) provides a suitable framework to study mathematically various mechanisms along these lines:

1. Modeling in (2.1);
2. Passing to the limit $\epsilon \to 0$ in (2.2);
3. Mathematical study of the limit problem;
4. Numerical comparison with (2.1).

This is the scheme of Chapter 10 in the case $d = 0$ (*i.e.* space-homogeneous, which can also be interpreted as a solution to (2.1) with homogeneous initial data). We study the behavior of solutions as $u = u^\epsilon$ solves an optimal control problem.

In [176] (which is Chapter 6) we stick to dimension $d = 1$ and study (2.2). The study is motivated by a remark in [29] about the influence of spatial variations of the population size. In the scalar model (of the form (2.2)), Barton and Turelli note that if the gene flow is modeled through an advection term, that is if $p$ satisfies

$$\partial_t p - \partial_{xx} p - 2\frac{\partial_x N}{N}\partial_x p = f(p), \tag{2.3}$$

where $N : \mathbb{R} \to \mathbb{R}_+$ is given (it is the population size) and $f$ is a bistable reaction term then for $\frac{\partial_x N}{N}$ large enough the traveling wave stops - the sign of its speed can even be reversed. We solve completely the case $\frac{\partial_x N}{N} = C\mathbb{1}_{[-L,L]}(x)$ for $C, L > 0$ to obtain Theorem 6.2. In substance, for fixed $L$ and large enough $C$ (or for fixed $C$ and large enough $L$) we show wave blocking. Otherwise the wave is merely delayed.

In Proposition 6.4 we describe the asymptotic behavior of $LC_*(L)$ as $L$ goes to 0, where $C_*(L)$ is the critical value of $C$ upon which blocking occurs. This result is interpreted as follows: along a propagation direction, if the population size increases too rapidly, or steadily on a long enough distance then the invasion stops. A jump in the population size is sufficient to block invasion if and only if its magnitude measured by $N_R/N_L$ (population size after the jump over population size before the jump) exceeds

$$[N]_{\text{crit}} := \Big(1 - \frac{\int_0^1 f(p)dp}{\int_0^\theta f(p)dp}\Big)^{1/4}.$$

This limit is relevant for application to insect populations, since the abundance can vary brutally depending on ground cover or water availability. Thus the gene flow term is able to model such spatial heterogeneity.

[4] The population must establish in a large enough area.

As for Theorem 6.1, it states that if $N$ depends in fact on $p$ the (2.3) has the same nature as before (without advection): it is a bistable reaction-diffusion equation. However the traveling wave speed sign can be reversed.

In [212] (which is Chapter 7) we focus on initial data. The problem consists in determining how to release *Wolbachia*-carrying mosquitoes at a given moment in time to ensure infection establishment, in the presence of uncertainty - and thus to ensure its dispersal since the two concepts are practically equivalent in this case.

The domain where mosquitoes are released is bounded. There, we first show that the success probability goes to 1 as the number of release points goes to $+\infty$ (Proposition 7.2). Then we build compactly supported reference profiles (sub-solutions in dimension 1 obtained by solving the stationary elliptic problem on bounded domains; negative energy profiles in dimension $> 1$) which are called "bubbles" or "propagules". If the infection profile, at any moment in time, stands above one of these reference profiles then the infection will maintain and invade the whole domain (Theorem 7.1).

Then we show that these sufficient invasion conditions (standing above a propagule) are very hard to reach with only one release point. Multiple release points induce an initial data as a sum of gaussian profiles, for which we provide analytical results. By doing so we quantify the invasion probability, and the results are illustrated by numerical simulations.

As a notable consequence, there exists an optimal size for the release domain. Even in the presence of uncertainty, an experimenter may seek for designing the area where the mosquitoes will be released in order to maximize the success probability. If releases are too grouped, the diffusion will dominate and the infection will never reach the critical spatial range to propagate. On the contrary if they are overly scattered then there is no place where the infection will reach the critical frequency to propagate.

From a monotone systems viewpoint, the boundary between the two basins of attraction contains simultaneously profiles with low values but extensive range, and profiles reaching high values on very localized ranges.

This work is connected to [93] where the authors quantified numerically the impact on population establishment success of two macroscopic observables of the initial data: fragmentation and abundance, in bistable reaction-diffusion. They conclude that fragmentation is rather detrimental although its effect depends on the abundance level. In Chapter 7 we recover similar questions, since the release domain area which maximizes the invasion success varies with abundance, and this size typically affects initial data fragmentation: larger domains favor fragmentation.

### 2.3.2 Ordinary differential systems

Most of the remainder of the thesis is concerned with systems of ordinary differential equations.

First, in Chapter 8 we study a very simple two-dimensional system coming from a compartmental model. The two remaining dimensions model the eggs and larvae populations. The interesting feature of the model is that the hatching rate $h \geq 0$ depends on larval density (the other parameters are positive constants):

$$\begin{cases} \dfrac{dE}{dt} = bL - (h(L) + d_E)E, \\ \dfrac{dL}{dt} = h(L)E - (cL + d_L)L. \end{cases} \tag{2.4}$$

Building on this simple hypothesis we prove that if $h' > 0$ (positive feedback) then system (2.4) is destabilized[5] and its solutions exhibit periodic oscillations. These oscillations occur in the form of a Hopf bifurcation along a parameter which describes the feedback magnitude, $h'(\overline{L})$ (where $(\overline{L}, \overline{E})$ is a stable positive steady state when $h'(\overline{L})$ is small enough, Theorem 8.2), or by means of a slow-fast analysis when the egg stock $\overline{E}$ is large and slowly varying (Theorem 8.1).

By highlighting the possibility (from a modeling point of view) that such oscillations can be caused by a positive feedback on hatching is a first step toward a possible explanation of some field observations. Some fluctuations indeed do not seem to correlate with environment variations (see [119], [141]). This study ought to be continued in order to describe more precisely (in more

[5] On the contrary, if $h' \leq 0$ then (2.4) is monotone cooperative and possesses 1 or 0 positive steady states. If it has a positive steady state then this one is globally asymptotically stable in $(\mathbb{R}_+^*)^2$, otherwise $(0, 0)$ is globally asymptotically stable. This is why we speak of "destabilization" under positive feedback.

complex compartmental models than (2.4)) the hatching dynamics in genus *Aedes*. Hatching is certainly a nonlinear phenomenon critical to understand these populations.

The next two chapters are concerned with practical problems raised by vector control methods. They appear as control problems for differential dynamical systems.

Chapter 9 is motivated by the use of the incompatible insect technique (IIT, see Section 3.3.1). By repeated releases of sterilizing males (*i.e.* males which make sterile the females they copulate with) the aim is to reduce or even eliminate a pest insect population (here, mosquitoes in genus *Aedes*). The model we build is deliberately simple but takes into account the eggs dynamics (which is thought to be very important for these species) and a hypothetical Allee effect[6] quantified by a parameter $\beta > 0$.

$$\begin{cases} \dfrac{dE}{dt} = bF(1 - \dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_E E\big(1 - e^{-\beta(M+M_i)}\big)\dfrac{M}{M+M_i} - \mu_F F, \\ \dfrac{dM_i}{dt} = u(t) - \mu_i M_i \end{cases} \tag{2.5}$$

Here, $E$ stands for all immature stages, $M$ is the male density and $F$ is the *fertile* females density (those which were inseminated by a compatible male). $M_i$ is the incompatible male density and $u(t)$ is the release flux.

Mathematically, the Allee effect makes the trivial (extinction) state $\mathbf{0}$ stable for (2.5) even with $u \equiv 0$. It is a new feature for such models, which provides a new insight into the natural questions associated with IIT (how large must the releases be, and how many of them are necessary), both analytical and geometrical: we are faced with the description of a separatrix, which is a co-dimension 1 sub-manifold.

Thanks to the entomological literature we define a range of values of the various parameters. We show that the time dynamics is bistable in general, and the stable positive steady state is denoted by $\mathbf{E}_+$. Some elementary (but useful) properties of the separatrix are stated in Proposition 9.2. In particular we show that above some (finite) given number of eggs or fertile females the population will always recover toward its wild state if the releases are stopped. The dynamical system on $(E, M, F, M_i)$ induced by (2.5) is monotone with respect to the cone $\mathcal{K}^o := \mathbb{R}^3_+ \times \mathbb{R}_-$. We show that the population is driven toward extinction under constant release of sterilizing males, provided the amplitude of the release is large enough ($M_i(t) \equiv M_i > M_i^{\text{crit}}$). Then we prove analytical bounds on the time it takes to reach the extinction basin, using sub- and super-solutions (Proposition 9.3). We obtain similar bounds for impulsive periodic releases with amplitude $\Lambda$ and period $\tau$ (Propositions 9.6 and 9.7). To this aim we show that $M_i(t)$ converges toward the $\tau$-periodization of $t \mapsto \frac{\Lambda e^{-\mu_i t}}{1-e^{-\mu_i \tau}}$. These bounds can be used to estimate analytically (as functions of the parameters) a number of releases sufficient to ensure extinction, for fixed $\Lambda$ and $\tau$. The detailed numerical study of a special case (isolated population of *Aedes polynesiensis*) illustrates the results and bolsters their practical interest.

Chapter 10 complements the previous chapter. It takes the control theory viewpoint on a similar problem: to steer a system of ordinary differential equations toward a stable equilibrium, starting from another stable equilibrium. The systems from Chapters 9 and 10 are both monotone with monotone control: releases of sterilizing males in the first place and of males-and-females *Wolbachia*-carrying mosquitoes here. The system from (10) is deduced from the homogeneous (2.1):

$$\begin{cases} \dfrac{dn_1}{dt} = b_1 n_1 (1 - s_h \dfrac{n_2}{n_1+n_2})(1 - \dfrac{n_1+n_2}{K}) - d_1 n_1, \quad n_1(0) = K(1 - \dfrac{d_1}{b_1}), \\ \dfrac{dn_2}{dt} = b_2 n_2 (1 - \dfrac{n_1+n_2}{K}) - d_2 n_2 + u, \quad n_2(0) = 0. \end{cases} \tag{2.6}$$

The controlled system (2.6) is complemented with a criterion describing the distance between the system's state at fixed time $T > 0$ and population replacement, that is the establishment of population 2 to the detriment of population 1:

$$J(u) := \frac{1}{2} n_1(T)^2 + \frac{1}{2}\big(K(1 - \frac{d_2}{b_2}) - n_2(T)\big)_+^2. \tag{2.7}$$

[6] When males are scarce, some females cannot get inseminated.

The constrained minimization problem associated with (2.6)-(2.7), for $u \in L^\infty(0,T)$ such that $0 \leq u \leq M$ and $\int_0^T u(t)dt \leq C$ for some $C, M > 0$ is denoted $(\mathcal{P}_{\text{full}})$.

We show rigorously (Proposition 10.2) that if fecundity is large ($b_i = b_i^0/\epsilon$, $i \in \{1,2\}$) and we take $\epsilon \to 0$ then $(\mathcal{P}_{\text{full}})$ converges toward a reduced problem $(\mathcal{P}_{\text{reduced}})$ defined by the minimization, with respect to the same control $u$ of the following functional:

$$J^0(u) := (1 - p(T))^2,$$

where $p$ solves

$$\frac{dp}{dt} = p(1-p)\frac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} + \frac{u}{K}\frac{b_1^0(1-p)(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p}, \quad p(0) = 0. \tag{2.8}$$

Equation (2.8) is interesting because it describes precisely how the simple (from a modeling point of view) control in (2.6) (releases of *Wolbachia*-carrying mosquitoes: $u(t)$ is the release flux) is transferred to a scalar control on the infection frequency.

As for problem $(\mathcal{P}_{\text{reduced}})$, it can be solved analytically (Theorem 10.1): for all $M > 0$ there exists $C^*(M) > 0$ (which can be expressed in terms of $b_i^0$, $d_i$, $s_h$ and $K$) such that if $C > C^*(M)$ the solution of $(\mathcal{P}_{\text{reduced}})$ is unique and equal to $u^* = M\mathbb{1}_{[0,C/M]}$ while if $C < C^*(M)$ this solution is unique and equal to $u^* = M\mathbb{1}_{[T-C/M,T]}$. If $C = C^*(M)$ then the set of minimizers is equal to $\{u_\lambda^* := M\mathbb{1}_{[\lambda,\lambda+C/M]},\ \lambda \in [0, T - C/M]\}$.

Combining both results (convergence and limit problem resolution) we state that if population replacement is possible, and if fecundity rates are large, then it is near-optimal to release all individuals at the beginning with the largest possible flux. The numerical study of problem $(\mathcal{P}_{\text{full}})$ shows however that for $\epsilon > 0$, the actual optimal strategies can differ greatly from this limit strategy.

Part III ends with Chapter 11, which is a theoretical contribution to the study of seasonal dynamics. The starting questions is: what are the effects of taking into account different seasons in the previous population dynamics? First we try to describe as precisely as possible the case when there are only two seasons, one being favorable and the other one unfavorable. We show (Theorem 11.2) that if the nonlinearities are concave then we can easily find sufficient conditions for a "sharp seasonal threshold property". By this we mean that the relative duration of the two seasons is a critical parameter: if it is below some threshold then the population extincts, and if it is above then the dynamics converges toward a unique periodic profile. This preliminary result should be put in perspective with control problems associated with vector control methods. In particular, the study of the periodic limit cycle may prove useful. The extension to other types of nonlinearities should also be discussed, with a particular emphasis on the seasonal alternation of bistable and (extinction) monostable dynamics.

### 2.3.3 Evolution equation for a structured population

Finally, Chapter 12 is concerned with a population structured by its phenotype. The population state is described by a positive measure on a phenotype space $\mathcal{P}$, typically $\mathcal{P} \subseteq \mathbb{R}^d$. The model is motivated by the burning issue of insecticide resistance dynamics in mosquito populations. In [200], Schechtman and Souza used an ordinary differential equations model to describe an asymmetry between the time it takes for resistance to appear in a population exposed to insecticide and the reversal time, that is the time it takes to disappear after the insecticide is no longer used. This result was obtained in the case of a genetic resistance caused by two successive mutations. The first one yields a good resistance level at a high fitness cost, and the second strengthens resistance while largely mitigating the fitness cost. Our goal is to describe that kind of asymmetry (or any other interesting time dynamics) for a continuously structured population with sexual reproduction. This is an important novelty compared with existing models, many of which dealing with resistance to treatment in bacteria or cells populations with clonal reproduction (see for instance [153]).

We propose a fairly general model and treat it only mathematically. We aim at designing tools to treat the sexual reproduction term, which is nonlinear and nonlocal, though 1-homogeneous. The equation reads

$$\partial_t n(t,x) = \frac{1}{\rho(t)} \iint_{\mathcal{P}^2} K(x,y,z) n(t,y) n(t,z) dy dz - R(x,\rho(t)) n(t,x), \quad \rho(t) = \int_{\mathcal{P}} n(t,x) dx. \tag{2.9}$$

Here, $K(x,y,z) \geq 0$ is the weighted distribution of the progeny from a female with trait $y$ mated with a male with trait $z$. $R(x,\rho)$ is the death rate of individuals with trait $x$ when the total population is equal to $\rho$. Such a model is justified only if the sex ratio is constant in time and also between different traits $x \in \mathcal{P}$. We also assume that $R(x,\cdot)$ is an increasing and unbounded function which models environment saturation. This allows to bound $\rho$ uniformly along forward orbits. First, we describe the singular properties of "imitative" models (this term comes from game theory, see [199]), where the offspring inherits exactly the trait of one of its parents. For these special models, a Lyapunov functional can be designed in some cases, allowing to prove concentration of the phenotypes toward the fittest trait (Theorem 12.5). Then we study the long time dynamics, when the mutation scale (measuring the distance to an imitative dynamics) becomes small. In this approach we construct a (classical) limit object in the form of a constrained Hamilton-Jacobi equation, which appears with new terms due to sexual reproduction (Theorem 12.4 and Section 12.6). Assuming for instance $\epsilon K_\epsilon(x,y,z) = B(y)\alpha(\frac{x-z}{\epsilon},y)$, where $\int \alpha(z',y)dz' \equiv 1$, *i.e.* the fecundity of a crossing between a $z$ male and a $y$ female only depends on $y$, then in the limit $\epsilon \to 0$ we get

$$\partial_t u(t,x) = \int B(y)q(t,y)\mathcal{L}[\alpha(\cdot\,,y)](\partial_x u(t,x))dy - R(x,\rho(t)), \quad \max_{x\in\mathbb{R}} u(t,x) \equiv 0,$$

where $\mathcal{L}[\alpha(\cdot\,,y)]$is the Laplace transform of $\alpha(\cdot\,,y)$ :

$$\mathcal{L}[a](p) := \int a(z)e^{-p\cdot z}dz,$$

and $q(t,y) = \lim_{\epsilon\to 0} n_\epsilon(t,y)/\rho_\epsilon(t)$ is the population distribution.

At this stage the study of the limit equation is incomplete. It is difficult due to time-dependent terms, $R(x,\rho(t))$ on the first hand and $q(t,y)$ on the other hand, which must be defined precisely by passing to the limit at the $\epsilon$ level equation. A first step has been made by obtaining uniform bounds in local total variation on $\rho_\epsilon(t)$ (Theorems 12.1, 12.2 and 12.3, for three particular cases). The extension of these results to more general reproduction kernels $K$ is discussed.

The problems of existence and uniqueness of steady state distributions for (2.9) are proposed as open problems in Section 13.4.

# Part I

# Context

# Chapter 3

# Mosquitoes and vector control

In this chapter, we start the context presentation by collecting important facts about vector mosquito species of genus *Aedes* (Diptera:Culicidae). These mosquitoes, in particular the species *Ae. aegypti* and *Ae. albopictus* (but also locally *Ae. polynesiensis*) are the main vectors of various arboviruses (dengue, chikungunya, zika). The population dynamics models developed during this thesis can apply to these species.

First we give a quick overview of what *Aedes* mosquitoes are, where they live and how they behave. Then, we define the vector-borne diseases under study and discuss important consequences for vector control. Finally, we describe in greater details two current approaches in vector control: population reduction by Sterile and/or Incompatible Insect Technique (SIT/IIT) and population replacement strategies.

## 3.1 Bio-ecology and monitoring of *Aedes* mosquitoes

*Aedes*[1] is traditionally the name of a "large genus comprising 931 species divided among 78 subgenera" (after [4]), whose first description is usually credited to Meigen [166]. The Culicidae taxonomy has been revised (since [108]), and the three insects of interest in this thesis, once belonging to *Aedes* genus, are now classified in *Stegomyia*[2] genus. Yet, they are widely known (in particular in tropical medicine) under their old names, which we retain throughout this work. The "new" *Stegomyia* genus comprises 128 species (in April 2018, [4]), among which *Aedes aegypti* (= *Stegomyia aegypti*) and *Aedes albopictus* (= *Stegomyia albopicta*) are major arbovirus vectors (see Section 3.2 for more details about vector-borne diseases).

Once a zoophilic and forest species from Africa [173], some *Ae. aegypti* populations adapted towards anthropophilic behavior, making it one of the most dangerous (for humans) mosquito species.

### 3.1.1 Life-cycle

The *Aedes* mosquitoes are **holometabolous** insects, which means that the larvae undergo a full metamorphosis to become adults. For this reason adults and juveniles have totally different ecology and behavior. Before moving to modeling works, it is important to keep in mind a clear notion of the mosquito's life-cycle. We reproduce below the overall description of the life-cycle in sub-family Culicinae (Diptera:Culicidae) that can be found in the review [76, Chapitre 11, pp. 251-253].

The life-cycle of Culicinae is divided into two phases: aquatic (egg, larva, pupa) and aerial (adults or "imago", both male and female). Adult females are usually considered to be inseminated by a single male, although multiple mating is possible (on this topic, see [180]). Roughly speaking, egg maturation takes about 3 days (but depends on temperature and varies among species) and a female can lay $40 - 80$ eggs per oviposition. A female *Ae. aegypti* or *Ae. albopictus* can split its eggs clutch between several breeding sites (see [48], [62]).

In particular for the three species under consideration, eggs can resist dessication and wait for several months before hatching. Eggs from *Ae. albopictus* are known to resist to low temperatures,

[1] The name *Aedes* comes from the greek word for "unpleasant" [2].

[2] Described in [210] citing [215] as "having scales completely covering the dorsal surface of the adult fly", whence this name from the greek words for "covered, roofed" and "fly".

allowing for the colonization of tempered areas, including Europe (see [104], and also [97] for a modeling work).

When stimulated, eggs **hatch**[3] and give rise to larvae, who feed on small particles in the water, and take from 3 days to several weeks (in particular in tempered areas) to develop fully and reach the pupa stage. Pupae still move in the water but do not feed anymore. This stage lasts $1-3$ days and leads to **emergence**, that is the beginning of the (aerial) adult stage.

Only females suck blood (on vertebrates), and usually take $2-5$ days between two blood meals (although one meal can consist in several bites if the insect is disturbed). It uses the proteins from the blood to maturate its eggs. Therefore, the ovipositions also occur every $2-5$ days. The mean lifespan for adults is estimated at a few weeks (for *Ae. albopictus* in laboratory conditions, 18 days for males and 30 days females according to [76, pp. 252-253]). In particular, a single female will perform several blood meals and ovipositions during its life, on average.

Adults can fly and their dispersal (depending on the environmental conditions like blood or breeding sites availability) is usually less than 1 km. Due to the excellent resistance of eggs, however, *Aedes* species have been transported between continents by human activity (see [165]).

### 3.1.2 Behaviors

The urban *Ae. aegypti* populations (for instance in Rio de Janeiro, Brazil, as studied in Chapter 8) is highly domesticated (see [76, pp. 259 - 260]) and lays eggs mostly in artificial containers (see for instance [119], [158]). It is **anthropophilic** (feeding preferentially if not only on humans), **endophagic** (feeding in houses) and **endophilic** (resting in houses). *Ae. aegypti* females bite during the day, with peak activity in the morning and late afternoon. The species is limited to tropical areas since it cannot resist low temperatures.

On the contrary, the range of *Ae. albopictus* (see [76, p. 260 - 261]) is much wider (in 2015, it was established in at least 20 European countries). It can use both natural and artificial containers and is overall **opportunistic**, feeding on humans as well as on many other vertebrate species. In particular in Rio de Janeiro, and also in any place where they are sympatric, *Ae. albopictus* and *Ae. aegypti* compete at the larval stage. In America, in Africa and in the Indian Ocean this interaction seems to be in favor of *Ae. albopictus*, while in South-East Asia the converse seems true.

According to [105, p. 8], in French Polynesia for *Ae. polynesiensis* the "aquatic stage lasts for $9-16$ days, adults can live up to $25-30$ days, laying around 5 egg batches of approximately 100 eggs each". *Ae. polynesiensis* is distributed throughout the eastern part of the South Pacific (see [105, p. 10]). Interestingly, this population shares with the *Ae. albopictus* population from Rio de Janeiro the ability to use both natural and artificial breeding sites. In the area under study in Chapter 9, coconuts (in particular rat-eaten ones) seem to be the main breeding site (see [105] and the references therein). This species is **exophilic**, opportunistic with a bias towards anthropophily.

### 3.1.3 Data acquisition

Acquiring data on mosquito populations is by no means an easy task. We gather here the four types of data collection that we have encountered during this thesis, and explain their main advantages and drawbacks.

First, **trapping counts** are widely used to measure at least relative variations in population abundance. The usual process relies on installing a trap in the field, and collecting on a regular basis the captured individuals (it can be adults but also eggs, or even preferentially gravid females, depending on the trap). Then, these specimens are identified. The data are typically counts of individuals captured in a trap at a given location, over time. Such data are used in [141] to try and estimate the relative abundance of *Ae. aegypti* in Rio de Janeiro during different seasons, and in [142] for surveillance quality assessment purposes. Such data are rather easy and cheap to collect, although it may require a large workforce in particular in remote locations or in the presence of many different species. They are rather good at estimating relative population abundance, when adequate traps are used in similar conditions. However, the trapping process itself is non-trivial (so that it needs to be modeled in order to interpret properly the data, see for instance [71, Chapters 3 and 4]), and in the long run the local population may adapt to the trap, inducing new biases.

[3] For further discussion of the hatching behavior of *Aedes* mosquitoes, see [78], [151], the modeling in [17] and [16] for Argentinian (=tempered) climate, the reference "life-table" model by Focks *et al.* [88] and the Chapter 8.

Second, once a lab colony has been established, **laboratory data** are commonly used to estimate vital parameters (see for instance [106] for *Ae. polynesiensis*) or vector competence (for instance in [34] for an *Ae. albopictus* population carrying *Wolbachia*, see Section 3.2 for a discussion on vector-borne diseases). When using a controlled environment with field-like conditions (like a large outdoor cage rather than a smaller one within an insectary), such experiments are called "semi-field" (as in [49]). These data are the easiest ones to collect (of course, only once the lab colony exists). However, one cannot expect that they inform precisely the field behavior, since the controlled conditions can never imitate perfectly the field ones, and over time (in spite of back-crossing), the lab population gradually digresses from the field one where it originated.

Comparing lab results (obtained from a lab colony) with similar experimental results obtained from field collected individuals (thanks to traps) is usually a good way to try and validate a hypothesis, as is done in [130] for the so-called "competition-longevity" hypothesis in *Ae. aegypti*.

Third, the **genetic data** can be used very efficiently, for instance among trapped individuals, to study the population's structure (see for instance [240]) and get much more detailed information than mere counts at each trap location. Such analyses are still expensive and can usually not be applied to a whole field collection, limiting the outcomes. However, the obtained data are extremely valuable, as was shown in [201] for confirming and quantifying barriers in the spread of introduced *Wolbachia* infection.

Finally, **mark-release recapture** (MRR) experiments rely on the release of marked (usually with powder) adults (or even pupae) from the lab into the field, with a trapping network around the release locations (see details and references in [71, Section 1.5.2]). This method was used for instance in [5] to estimate comparative dispersal of *Ae. aegypti* and *Ae. albopictus*, and in [229] to estimate population abundance of *Ae. aegypti*. It can also be used to estimate adult dispersal. The process itself is not very expensive (although it requires good logistics) but the data are not necessarily easy to use, depending a lot on the recapture rate. In addition, the behavior of a large number of adults released at a single (or at a few locations) after being raised in laboratory cannot be expected to mimic closely the natural behavior of a wild adult. Still, at this stage MRR seems hard to overcome as a standard tool to improve our knowledge about mosquito behavior in the field.

## 3.2 Vector-borne diseases

### 3.2.1 Vectors and vector-borne diseases

The three mosquito species presented above (*Ae. aegypti*, *Ae. albopictus* and *Ae. polynesiensis*) are of interest in medical entomology because they are *vectors*. After [76, Chapitre 2, p.44], we define:

**Definition 3.1.** *A **vector** is an arthropod which actively transmits an infectious agent.*

*In medical entomology, a **vector-borne disease** (VBD) is any infectious human disease whose agent (parasite, virus, bacterium etc.) can be transmitted by a vector.*

Here, the vector is a blood-sucking arthropod (a mosquito of genus *Aedes*) and the infectious agent is a virus (for instance, dengue).

*Ae. aegypti* and *Ae. albopictus* are main vectors for dengue, which is currently the most severe mosquito-transmitted arboviral disease worldwide (see [32] where the authors estimate that 390 million infections occur annually, of which 96 million are manifest). Other arboviral diseases have emerged in the past years (in particular zika), and the overall burden of these for public health is still hard to assess (see [184] for a recent account of the (re-)emergence of dengue, chikungunya and zika). In 2016 however, an analysis [204] attempted to estimate the economic burden of dengue, which is expected to be huge (in particular, the cited paper estimates that there have been 13586 fatal cases in 2013, with 95% uncertainty interval 4200-35700).

The main focus of the thesis is the mathematical mosquito population modeling, and therefore we do not delve any further into epidemiology and vector-borne diseases modeling, and refer the interested reader to [76, pp. 271 - 288] and the references therein for a more thorough discussion of infectious diseases transmitted by Culicinae mosquitoes from a public health point of view.

### 3.2.2 Vector competence and vectorial capacity

Only adult female mosquitoes bite, at least once per **gonotrophic cycle**: the female needs a blood meal to maturate eggs, and therefore every oviposition is preceded (by a few days in general) by a blood meal on a vertebrate. Each blood meal is an opportunity for the infectious agent to circulate either from the vertebrate's blood to the arthropod's gut or from the arthropod's saliva to the vertebrate's blood. (For more details on *Aedes* mosquitoes, see Section 3.1)

The transmission of the virus from mosquito to human being is done through the saliva, during a bite. It is therefore called *biological*, since biological (and not merely mechanical) processes are required for the virus to move from the infected blood meal to the mosquito's salivary glands.

Following [76, Chapitre 2, p. 52 et seq.], the arbovirus transmission is divided in three main steps:

**The virus infects the vector** during a blood meal on an infected vertebrate with sufficiently high viremia;

**The virus multiplies in the vector** and manages to reach the salivary glands;

**The virus leaves the vector** in its saliva.

Note that the first step occurs only if the vector effectively feeds on infected vertebrate, while the second and the third step require a good match between the virus and the vector. After these steps, the vector is qualified as **infective**, which means that any of its following bites can potentially infect a vertebrate.

**Definition 3.2.** *The time required to complete the three steps of arbovirus transmission is called the* ***extrinsic incubation period*** *(EIP).*

*The ability of a vector population to get infective for a given infectious agent is called* ***vector competence****. It can be quantified as the frequency of vector individuals which get infective after a blood meal on an infected vertebrate (cf. [77],[152]).*

Vector competence must be handled with care since it may vary a lot depending on the viremia of the vertebrate, the population, or even slight mutations of the virus (*cf.* [228]).

In principle, the shorter the EIP and the larger the vector competence, the better the VBD can circulate. However, other factors come into play, mostly from ecology: the vector and the vertebrate species must have frequent and strong enough contact for the VBD to circulate. This leads to the concept of vectorial capacity:

**Definition 3.3** (After [77])**.** *The* ***vectorial capacity*** *quantifies the ability of a vector population, in a given environment, to transmit a given virus to a given human population. It is the daily rate at which future inoculations arise from a currently infective case.*

Mathematically, vectorial capacity has been defined from a basic case reproduction number in an epidemiological model (see [66] for the definition) for malaria transmission (see [157], [94] and [77]). According to [77], this approach has proved very useful. Although the absolute threshold computation is virtually impossible due to data availability for each parameter, it allows for effective comparison either between diseases, areas or species and gives valuable insight into vector control. Dye [77] gives the following formula for vectorial capacity:

$$VC = \frac{ma^2 bp^{\tau_{EIP}}}{-\log(p)},$$

where $\tau_{EIP}$ is the duration of the extrinsic incubation period, $a$ is the biting rate (number of bites on humans per female per day), $m$ is the relative abundance (number of active females per human), $p$ is the daily survival rate of females ($-\log(p)$ is expressed in days), and $b \in [0, 1]$ is the vector competence (as defined above).

### 3.2.3 Vector control

**Vector control** (VC) measures are human intervention with two objectives (see [76, Chapitre 5]):

- protect individuals from infectious bites,

- prevent or reduce the circulation of VBD within a community.

Key parameters of vectorial capacity are natural targets for VC: vector density and contact with host (reduce $m$, $a$) and vector lifespan (reduce $p$). According to the classification in [22], vector control measures include environmental fight (breeding site destruction), mechanical fight (trapping), chemical fight (use of chemical or bacterial insecticide), biological fight (predator introduction or *Wolbachia* replacement strategy) and genetic fight (RIDL, gene drive, sterile insect technique, see [7]). Apart from gene drive and *Wolbachia* replacement strategies, these measures focus on reducing $m$ and/or $p$.

Recently, **population replacement strategies** using the natural bacteria *Wolbachia* have been developed for *Ae. aegypti* (see the 2008 review [38] and the series of papers during the period 2009-2014, in particular [172], [232], [118] and [117]). Their primary target is the vector competence $b$, which is a new feature for vector control. This is beneficial at several levels (see [164] for a detailed account). Being specific to the vector species, it does not raise the ecological issues that chemical control does. The reduction of vector competence for several viruses in *Ae. aegypti*, if stable in time, could practically suppress disease circulation and the need for further vector control measures in some areas. (See Section 3.3.2 for more details).

Meanwhile, there has been a renewed interest for **population reduction or elimination programs** relying on releases of sterile (or sterilizing) males, as detailed below in Section 3.3.1.

## 3.3 Two vector control methods by releases

### 3.3.1 Population reduction by SIT/IIT

The Chapter 9 introduces a new model for population reduction by releases of sterilizing males, developed to address an experimental setting in the French Polynesian atoll of Tetiaroa.

Pest management through the **Sterile Insect Technique** (SIT) has a long history. SIT is a promising technique that has been first studied by E. Knipling and collaborators, and first experimented successfully in the early 1950s by nearly eradicating screw-worm population in Florida (see the biographical note [6]). Since then, SIT has been applied on different pests and vectors, like fruit fly or mosquito. The classical SIT relies on the mass releases of males sterilized by ionizing radiation. The released sterile males transfer their sterile sperm to wild females, which results in a progressive reduction of the target population. For mosquito control in particular, SIT has been adapted using *Wolbachia*. *Wolbachia* is a bacterium that infects many Arthropods, and among them some mosquito species in nature. It was discovered in 1924 by Hertig and Wolbach [110]. Since then, various interesting features of these bacteria have been unveiled in many arthropod species, many of which are summarized in [233]. One of these properties is particularly useful for the control of *Aedes* populations: the cytoplasmic incompatibility (CI) [206]. CI can be used for two control strategies: Incompatible Insect Technique (IIT) or population replacement (see Section 3.3.2 for the latter).

IIT relies on the fact that the sperm of males carrying a CI-inducing *Wolbachia* strain is altered by the bacterium so that it can no longer successfully fertilize eggs from females who do not carry this strain of *Wolbachia*. This can result in a progressive reduction of the target population when incompatible males are released. Prior experiments [179] have shown the potential effectiveness of this method for lab and field *Ae. polynesiensis* populations. One limitation of IIT with respect to SIT is that in case of accidental release of females, the introduced *Wolbachia* population could establish in the field, which is not the case for SIT where the irradiation dose also sterilizes females, making sexing errors much less risky. Still, once a lab colony is established, IIT does not require costly equipments.

As was pointed out in [144], SIT/IIT techniques, potentially combining CI-inducing *Wolbachia* and irradiation (until the sexing is sufficiently accurate), are now considered *again* as promising tools for vector control.

### 3.3.2 Population replacement strategies using *Wolbachia*

Chapters 6, 7, 10 are direclty concerned with the modeling of population replacement strategies using *Wolbachia*.

This technique was published in 2011 both for caged populations [232] and field establishment [118]. It originates from the discovery of virus replication blocking phenotype of *Wolbachia*

strain wMelPop in *Ae. aegypti* [172] (known as **pathogen interference** or PI), for dengue, zika and chikungunya viruses.

When sufficiently many males and females carrying *Wolbachia* are released in a susceptible population, the CI phenotype will tend to favor the introduced *Wolbachia*-carrying population, and since the bacterium is maternally inherited, a long-lasting infection will establish in the field population. In other words, there can be a population replacement by *Wolbachia*-carrying mosquitoes, and this is why this technique is sometimes qualified as *self-sustaining* (as opposed to *self-limiting* ones such as SIT). Once the infection is established, the PI phenotype effectively limits the disease circulation (as explained in [118]).

It has been stated in [127] that statistical studies based on the data from the successful population replacement trial in Cairns, Australia have confirmed the local sustainability of this method (see [202]).

It is worth emphasizing that unlike any other VC method, population replacement strategies do not harm either the ecosystem or the mosquito population in itself, in the sense that the main effect of carrying *Wolbachia* in this context is the PI phenotype (this depends of course on the *Wolbachia* strain to be used, and of the species).

At the end of this chapter, we also underline the ongoing development of *Wolbachia*-based bioengineering methods for pest management in general, which may be a source for new and alternative vector controls techniques, as suggested for instance by [12]. The need for further mathematical modeling and optimization of these techniques may therefore be growing in the years to come, as innovative pest management concepts become real.

# Chapter 4

# Population dynamics modeling

In this chapter, we introduce the mathematical tools used to model population dynamics in this thesis, with a particular emphasis on the biological interpretation of the mathematical properties. The history of population dynamics modeling is not developed in this chapter and we simply refer readers interested in this topic to the book [18]. A clear presentation of population dynamics modeling with relevant references can also be found in the PhD thesis of Claire Dufourd [71, Section 1.4].

First we give notations, definitions and our biological motivation. Then we gather some general results on differential equations, on reaction-diffusion equations and also some other results used in the following chapters.

## 4.1 Notations, framework and motivation

### 4.1.1 Notations

We use the following notations without reference or re-definition:

- $:=$ stands for a definition,
- $\mathbb{R}$ is the field of real numbers, $\mathbb{R}_+^*$ is the set of positive real numbers, $\mathbb{R}_+ := \mathbb{R} \cup \{0\}$, $\mathbb{Z}$ is the ring of integers, $\mathbb{Z}_{>0}$ is the set of positive integers and $\mathbb{Z}_{\geq 0} := \mathbb{Z}_{>0} \cup \{0\}$,
- for $x \in \mathbb{R}$, the notation $\lfloor x \rfloor$ stands for the largest integer $n \in \mathbb{Z}$ such that $n \leq x$,
- for $n, m \in \mathbb{Z}_{>0}$, $M_{n,m}(\mathbb{R})$ is the set of matrices with $n \times m$ entries in $\mathbb{R}$: $n$ lines and $m$ columns. The algebra of square matrices of size $n$ is denoted $M_n(\mathbb{R}) := M_{n,n}(\mathbb{R})$,
- for $n \in \mathbb{Z}_{>0}$, $\mathbf{I}_n \in M_n(\mathbb{R})$ is the identity matrix in dimension $n$ (the subscript $n$ is dropped when unambiguous),
- the group of invertible matrices in $M_n(\mathbb{R})$ is denoted by $\mathrm{GL}_n(\mathbb{R})$,
- the largest real part (resp. modulus) of an eigenvalue of $A \in M_n(\mathbb{R})$ is denoted by $\mu(A)$ (resp. $\rho(A)$).
- the euclidean scalar product in $\mathbb{R}^d$ is denoted by $\langle \cdot, \cdot \rangle$. The same notation is sometimes used for the dual evaluation: given a continuous linear form $f$ on a Banach space $E$, and $x \in E$, we write $\langle f, x \rangle = f(x)$.

Let $d \in \mathbb{Z}_{>0}$ and $\Omega \subset \mathbb{R}^d$ be an open set.

- If $f : \Omega \to \mathbb{R}^{d'}$ is Fréchet-differentiable, we denote the differential of $f$ at $x \in \Omega$ (which is a matrix in $M_{d,d'}(\mathbb{R})$) either by $DF_x$, $\nabla F_x$ or $(\partial_{x_i} f(x))_{1 \leq i \leq d}$, dropping the subscript $x$ when convenient. The two notations $\partial_{x_i}$ and $\frac{\partial}{\partial x_i}$ $(1 \leq i \leq d)$ are used equally. The notation $D$ is also used when $\Omega$ is an open subset of a normed real vector space.
- A multi-index is $\alpha \in \mathbb{Z}_{\geq 0}^d$, and the multi-derivative of a function $f : \Omega \to \mathbb{R}$ is denoted by $D^\alpha f(x) := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f$. The length of a multi-index is denoted by $|\alpha| := \sum_{i=1}^d \alpha_i$.

- For $m \in \mathbb{Z}_{>0}$ and $p \in [1, \infty)$, we use the Sobolev spaces (see [40])

$$L^p(\Omega) := \{f : \Omega \to \mathbb{R} \text{ measurable}, \int_\Omega |f(x)|^p dx < +\infty\},$$

$$W^{m,p}(\Omega) := \{f : \Omega \to \mathbb{R},\ D^\alpha f \in L^p(\Omega) \text{ for all multi-index } \alpha, |\alpha| \leq m\}.$$

  We also define $H^m(\Omega) := W^{m,2}(\Omega)$, which is a Hilbert space. For $p \in [1, \infty)$ and $f \in L^p(\Omega)$, the $L^p$-norm of $f$ is $\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p dx\right)^{1/p}$. The $L^\infty$ space and the corresponding norm are defined by

$$L^\infty(\Omega) = \{f : \Omega \to \mathbb{R} \text{ measurable}, \ \exists C > 0, |f(x)| \leq C \text{ for a.e. } x \in \Omega\},$$

  and $\|f\|_{L^\infty} := \inf\{C > 0, |f(x)| \leq C \text{ for a.e. } x \in \Omega\}$. For $p \in [1, \infty]$, we use equally the notations $\|\cdot\|_{L^p(\Omega)}$ and $\|\cdot\|_p$, when convenient.

- If $X$ is a complete metric space, $\mathcal{M}(X)$ denotes the set of Radon measures on $X$, $\mathcal{M}_+(X)$ the pointed cone of positive measures and $\mathcal{M}_+^1(X) \subset \mathcal{M}_+(X)$ the set of probability measures.

### 4.1.2 Mathematical framework

From an abstract viewpoint, the population state at time $t \in \mathbb{R}$ is described by a quantity $n(t) \in \mathcal{K}$, where $\mathcal{K}$ is a cone and a subset of a normed vector space $\mathcal{X}$. In order to get practical, let us describe in brief the state spaces $\mathcal{X}$ used thereafter.

- when using differential systems, we consider that the population state at a given time is defined as a finite number of values: the headcount of each group. Hence $\mathcal{X} = \mathbb{R}^{N_d}$ (where $N_d \in \mathbb{Z}_{>0}$ is the number of components) and $\mathcal{K} = \mathbb{R}_+^{N_d}$.
- when dealing with reaction-diffusion equations, we take into account the spatial dispersal of the population, so for each group in the population the state is a locally integrable (and non-negative) function of the space variable. We will typically consider $\mathcal{X} = \left(H^1(\Omega)\right)^{N_d}$ (where $\Omega \subset \mathbb{R}^d$ is a domain and $d \in \mathbb{Z}_{>0}$ is the physical space dimension, typically $d \in \{1, 2\}$), and $\mathcal{K} = \{u \in H^1(\Omega),\ u \geq 0 \text{ almost everywhere}\}^{N_d}$.
- for structured populations, the number of groups within the population no longer needs to be finite; we rather group individuals sharing the same phenotype, which may vary continuously (for instance, size or age). Formally, there is a base phenotype space $\mathcal{P}$ (a complete metric space, typically $\mathcal{P} = \mathbb{R}^d$ for some $d \in \mathbb{Z}_{>0}$) and $\mathcal{X} = \mathcal{M}(\mathcal{P})$, $\mathcal{K} = \mathcal{M}_+(\mathcal{P})$.

Deterministic population dynamics considered here take the form of a differential equation on the Banach space $\mathcal{X}$, defined by some function $\mathcal{F} : \mathbb{R} \times \mathcal{X} \to \mathcal{X}$ with

$$\frac{dn}{dt} = \mathcal{F}(t, n(t)). \tag{4.1}$$

In all cases, we assume that all orbits starting in $\mathcal{K}$ remain in $\mathcal{K}$.

**Definition 4.1** (Positive differential dynamical systems)**.** *The dynamics of (4.1) is called $\mathcal{K}$-positive (or "positive") if*

$\forall t < t'$ *and* $n^0 \in \mathcal{K}$ *such that the solution* $n$ *to (4.1) with* $n(t) = n^0$ *is defined on* $[t, t']$, $n(t') \in \mathcal{K}$.

Positive systems are suitable for population dynamics modeling because no non-positive state $n \in \mathcal{X} \backslash \mathcal{K}$ could be interpreted as a population: therefore it is a bare requirement of the models that they always represent a population as a positive quantity.

We introduce the now classical comparison notations for the partial order on $\mathcal{X}$ induced by a cone $\mathcal{K}^o$ (widely used in monotone systems theory, see [115])

**Definition 4.2** (Partial order induced by a cone)**.** *For $A, B \in \mathcal{X}$ and $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$:*

- $A \leq_{\mathcal{K}^o} B$ *if and only if* $B - A \in \mathcal{K}^o$,
- $A <_{\mathcal{K}^o} B$ *if and only if* $A \leq_{\mathcal{K}^o} B$ *and* $A \neq B$,

- $A \ll_{\mathcal{K}^o} B$ *if and only if* $B - A \in \mathring{\mathcal{K}^o}$,
- $\mathcal{A}\mathcal{R}_{\mathcal{K}^o}\mathcal{B}$ *if and only if* $\forall (A,B) \in \mathcal{A}\times\mathcal{B}$, $A\mathcal{R}_{\mathcal{K}^o}B$ *(for* $\mathcal{R} \in \{\leq, <, \ll\}$*).*

The subscript $\mathcal{K}^o$ is dropped hereafter when it is obvious from the context. Many of the dynamical systems studied in this thesis are *monotone* in the sense of [114]:

**Definition 4.3** (Monotone differential dynamical systems)**.** *The dynamics induced by* (4.1) *is called* $\mathcal{K}^o$*-monotone (or simply "monotone") if for all solutions* $n_1, n_2$ *of* (4.1) *and* $t < t'$*, if* $n_1(t) \leq_{\mathcal{K}^o} n_2(t)$ *then* $n_1(t') \leq_{\mathcal{K}^o} n_2(t')$*. It is called "strongly monotone" if*

$$\forall t < t', \quad n_1(t) <_{\mathcal{K}^o} n_2(t) \implies n_1(t') \ll_{\mathcal{K}^o} n_2(t').$$

*It is called "strongly order-preserving" (SOP*[1]*) if for all* $n_1^0 <_{\mathcal{K}^o} n_2^0$ *there exists neighborhoods* $U$ *and* $V$ *of* $n_1^0$ *and* $n_2^0$ *and* $t_0 \geq 0$ *such that for all* $t > t_0$*, the images of* $U$ *and* $V$ *by the semiflow* $\phi^t$ *of* (4.1) *satisfy* $\phi^t(U) \leq_{\mathcal{K}^o} \phi^t(V)$*.*

Monotonicity means that the partial order induced by $\mathcal{K}^o$ is preserved by the time-dynamics. For monotone systems, "more" population at a given time will always yield "still more" population at further times. Note that the order-inducing cone $\mathcal{K}^o$ needs not be equal to the positive cone $\mathcal{K}$. For competitive differential systems we will typically have $N_d = N_d^+ + N_d^-$ for some $N_d^\pm \in \mathbb{Z}_{>0}$ and $\mathcal{K}^o = \mathbb{R}_+^{N_d^+} \times \mathbb{R}_-^{N_d^-}$.

### 4.1.3 Interpretation and motivation

The above three base state spaces $\mathcal{X}$ correspond to different modeling levels. **Differential equations** ($\mathcal{X} = \mathbb{R}^{N_d}$) are sometimes termed "mean-field" approaches (see [74]). In this approach, individuals within the population are classified into a finite number of categories, depending on macroscopic features such as "being an adult", "being a male" or "being an egg laid at a given oviposition site". This approach can be very efficient in reducing the system's dimensionality. However, it erases the individual variations, and therefore in many cases it neglects parameters (such as the age or the amount of resources left in a given egg) that may reveal crucial in determining the future individual behavior (for instance, mating or hatching). **Reaction-diffusion equations** (see, for instance, [187]) are used as simple generalizations of the mean-field approach to space-varying population densities. They take into account the spatial parameter, which is usually of tremendous importance for any interaction: individuals that are further apart are much less likely to interact with each other. **Structured populations** (see, for instance, [186, Chapter 1]) offer a rich formalism in which individuals can be classified using much finer feature than the macroscopic ones from the mean-field approach. In particular, continuous phenotypes are possible and have proved very relevant to model age, size, the expression level of some protein, or any combination of features.

The modeling philosophy in this thesis aims at defining low-dimensional or simple models to represent a very complex and little-known natural population of insects (see Chapter 3). Therefore, we stress that the models we propose and study mathematically are only able to answer practical questions *assuming that* all neglected features or parameters play a minor role. The bright side of this approach is that our results are *analytical*, and rely on mathematical proofs. They do not depend on particular numerical or biological experiments and are, in this sense, perfectly reproducible, while the underlying assumptions are clearly stated. On the downside, these results are in some sense too *rigid*, due to the structural properties of the models: a deterministic outcome is usually irrelevant in practice since many factors are stochastic. However, this determinism in itself is very useful when the models are used properly, that is to study qualitatively and also quantitatively the *relative* influence of various parameters or mechanisms. Moreover, one should not forget that we model population counts (which are in fact integers) by continuous functions.

The validation of such models is a tricky point, and has been little explored during this thesis. It is worth repeating that the expected outcomes are not population dynamics prediction indeed, but rather proofs of qualitative facts such as "these mechanisms can (or cannot) explain *by themselves* these observations", or answers to comparative questions. In addition to these qualitative outcomes, some quantitative outputs are still relevant, mostly in terms of *scales* or of key parameters identification.

[1] In spite of its technical statement, the SOP property boils down to an irreducibility condition on the Jacobian of the right-hand side for ordinary differential equations, see for instance [209].

When building new models (in Chapters 5, 8, 9 and 12), the starting point has always been a double observation:

- that some experimental data where consistently suggesting non-linear effects, with reasonable biological hypotheses clearly formulated on the possibly involved mechanisms;
- that no existing mathematical model had been, up to our knowledge, proved satisfactory in reproducing the observed output using only the hypothesized mechanisms.

Some details on the motivation and the modeling process are given below (see also the introduction of the cited chapters for academic background on these topics).

The spreading of *Wolbachia* in an insect population was pretty well understood (see [29]) using a scalar reaction-diffusion equation on the proportion of infected individuals (this model is studied in Chapters 6 and 7); however, the underlying assumption that the dynamics only depends on the proportion and not on the total population size was not justified (except assuming that the population size is infinite). We propose in Chapter 5 a derivation of the scalar model from a two-populations model, in a small-parameter regime (= singular limit) corresponding to large fecundity.

Likewise, the oscillations in the trapping data were thought to be well-explained by climate variations (in particular, rainfall and temperature), but field data from Rio de Janeiro (see [119]) suggested that this was not enough, and some underlying oscillating behavior was in place in the local *Aedes aegypti* population. Researchers from Fundação Oswaldo Cruz hypothesized that a previously identified egg hatching stimulation by larvae (see [78]) could account for these oscillations. We introduced with them and study mathematically in Chapter 8 a simple non-linear model to try and confirm this assumption from a modeling point of view.

A recent experiment in a small island in French Polynesia led by the Institut Louis Malardé (ILM) showed a very good success for population elimination using an Incompatible Insect Technique (IIT). The existing mathematical models for SIT (Sterile Insect Technique, see Section 3.3.1), which can apply in this context, all assumed that the population extinction state was unstable in the absence of sterile male releases. In other words, as soon as the control is stopped, in these models, the insect population recovers towards its initial state. In Chapter 9, we develop and study a model incorporating the stability of extinction (with a tunable basin of attraction) to draw consequences and get a new insight into this problem.

Finally, Chapter 12 is a preliminary attempt at involving sexual reproduction into resistance to insecticide pattern formation, where the resistance level (= phenotypical trait) can vary continuously (which makes sense if it is associated with a protein expression level, for instance), can affect the fecundity (and not only the mortality, contrary to most existing models) and where the crossing between two individuals with known resistance levels determines the offspring trait distribution.

## 4.2 Differential equations

In Part III, the models under study are differential equations, since the population state is described by a finite number of (non-negative) real values, which vary continuously.

### 4.2.1 Monotone bistable differential systems

In Chapters 9 and 10, the systems (respectively in $\mathbb{R}^4$ and in $\mathbb{R}^2$) are monotone and bistable. We collect here some useful facts from the monotone dynamical systems theory concerning monotone bistable differential systems with linear control, in any dimension. The state is $n \in \mathbb{R}^d$ ($d \geq 1$), $f : \mathbb{R}^d \to \mathbb{R}^d$ is smooth, $\iota \in \mathbb{R}^d$ is a fixed unitary vector and $u : \mathbb{R}_+ \to \mathbb{R}_+$ is a control term. Let $(e_i)_{1 \leq i \leq d}$ denote the canonical basis of $\mathbb{R}^d$.

Let us consider

$$\dot{n} = f(n) + \iota u, n(0) = n^0. \tag{4.2}$$

After [115], we assume that for all $i, j \in [\![1, d]\!]$ the following holds:

**Positivity.** $f(n) \cdot e_i \geq 0$ if $n_i = 0$,

**Forward boundedness.** There exists a compact subset $X_0$ of $\mathbb{R}^d_+$ stable for (4.2),

**Sign stability.** $\sigma_{i,j} := sgn\left(\frac{\partial f_i}{\partial x_j}\right)$ does not change sign on $\mathring{X}_0$,

**Sign symmetry.** $\sigma_{i,j}\sigma_{j,i} > 0$,

**Sign consistency.** the signed incidence graph associated with the Jacobian matrix $Df(x)$, that is the graph with undirected edge joining vertices $i$ and $j$ if there exists $x, y \in X_0$ $\frac{\partial f_i}{\partial x_j}(x) + \frac{\partial f_j}{\partial x_j}(y) \neq 0$, the edge being given the sign of this sum, has the property that every loop has an even number of negative signs (negative feedbacks).

Under these assumptions, (4.2) is positive, forward bounded, and there exists a unique orthant (up to the sign) $\mathcal{K}^o$ of $\mathbb{R}^d$ such that it is $\mathcal{K}^o$-monotone on $X_0$. We assume in addition that the control itself is monotone:

**Control monotonicity.** $\iota \in -\mathcal{K}^o$.

**Proposition 4.1.** *Under these assumptions, if $u_1, u_2 : \mathbb{R}_+ \to \mathbb{R}_+$ are bounded in $L^1(\mathbb{R}_+)$ with compact support and satisfy $u_1 \leq u_2$, the semiflow $(\phi_i^t)_{t\geq 0}$ defined by (4.2) with $u = u_i$ ($i \in \{1, 2\}$) are monotone, that is for all $t > 0$ and $n^1 \leq_{\mathcal{K}^o} n^2$, $\phi_1^t(n^1) \leq_{\mathcal{K}^o} \phi_2^t(n^2)$.*

*Proof.* This follows from Kamke's theorem (see [114, Section 3.1]). □

We address the specific context of bistability:

**Bistability.** $X_0$ contains exactly two hyperbolic stable steady states denoted $\mathbf{0}$ and $\mathbf{E}_+$, such that $f(\mathbf{0}) = f(\mathbf{E}_+) = 0$ and $\mathbf{0} \ll_{\mathcal{K}^o} \mathbf{E}_+$.

**Proposition 4.2.** *The closed order interval $[\mathbf{0}, \mathbf{E}_+]$ is a global attractor.*

*Proof.* By [114, Theorem 3.14], the set of points converging to $\mathbf{0}$ or $\mathbf{E}_+$ is dense in $X_0$. In addition, the basins of attraction of $\mathbf{0}$ and $\mathbf{E}_+$ are $p$-convex (that is, order-convex, in the sense that if $u < v$ then the set contains the line segment spanned by $u$ and $v$). Using the comparison principle, the result follows immediately. □

Let us define the basins of attraction of these steady states and the separatrix:

$$\Sigma_+ := \{n^0 \in X_0,\ n(t) \xrightarrow[t\to+\infty]{} \mathbf{E}_+\}, \tag{4.3}$$

$$\Sigma_- := \{n^0 \in X_0,\ n(t) \xrightarrow[t\to+\infty]{} \mathbf{0}\}, \tag{4.4}$$

$$\Sigma := X_0 \backslash \left(\Sigma_- \cup \Sigma_+\right). \tag{4.5}$$

**Proposition 4.3.** *The basins of attraction $\Sigma_\pm$ are open. Let $x \in X_0$. If there exists $y \in \Sigma_+$ such that $y \leq_{\mathcal{K}^o} x$ then $x \in \Sigma_+$. Similarly, if there exists $y \in \Sigma_-$ such that $y \geq_{\mathcal{K}^o} x$ then $x \in \Sigma_-$. The union $\Sigma_+ \cup \Sigma_-$ is dense in $X_0$. The separatrix $\Sigma$ contains at least one (unstable) steady state. It contains no pair of points related by $\ll_{\mathcal{K}^o}$. It is in fact a $d-1$ dimensional submanifold.*

*Proof.* This result is in fact a consequence of the classical properties of SOP semiflows collected in [114]. We focus on the proof of the last point. Let $\mathbb{S}_+^{d-1}$ denote the intersection of the $(d-1)$-dimensional sphere (embedded in $\mathbb{R}^d$) with the order orthant $\mathcal{K}^o$. Then we can build a diffeomorphism $\psi : \Sigma \to \mathbb{S}_+^{d-1}$ by associating each point $v \in \Sigma$ to the unique $w \in \mathbb{S}_+^{d-1}$ such that $v - \mathbf{0} \in \mathbb{R}_+ w$. Existence and uniqueness are straightforward since $v > \mathbf{0}$. The inverse definition comes from the comparison principle: in any given direction $w \in \mathbb{S}_+^{d-1}$, there is at most one "separatrix value" $\lambda(w)$ such that $\psi^{-1}(w) := \mathbf{0} + \lambda(w)w \in \Sigma$ since $\Sigma$ cannot contain two ordered points. Regularity of $\psi^{-1}$ and its inverse comes from smoothness of $f$ and the facts that the basins $\Sigma_\pm$ are open and their reunion is dense. □

### 4.2.2 Slow-fast dynamics

Chapters 8 and 10 provide two different examples of two-dimensional slow-fast dynamics. In the former the small parameter is due to the assumption that the egg population is large and has slow dynamics compared with the larvae population, while in the latter it comes from the assumption that the fecundity rate is large compared with mortality rate. In both cases, the outcome is the simplification of the dynamics as the small parameter goes to 0.

A prototypical slow-fast system is given by:

$$\begin{cases} \frac{du_\epsilon}{dt} = f(u_\epsilon, v_\epsilon), \\ \epsilon \frac{dv_\epsilon}{dt} = g(u_\epsilon, v_\epsilon), \\ u_\epsilon(0) = u^0, \quad v_\epsilon(0) = v^0, \end{cases} \tag{4.6}$$

where $f$ and $g$ satisfy suitable assumptions. In fact, $f$ and $g$ can be replaced by (for instance) uniformly converging nets $(f_\epsilon)_\epsilon$ and $(g_\epsilon)_\epsilon$ (and also $u^0$ and $v^0$ by $u^0_\epsilon$ and $v^0_\epsilon$) under additional assumptions. In [217], Tikhonov proved asymptotic properties for non-autonomous systems of the form (4.6), and also further extensions. We state a simple result (the hypotheses are far from being optimal) to get the gist of these properties:

**Proposition 4.4** (Tikhonov). *In (4.6), assume that $u_\epsilon, v_\epsilon$ are defined on $[0, +\infty)$ and uniformly bounded by $K_u, K_v > 0$ (with respect to $\epsilon$), and that for all $u \in [-K_u, K_u]$, there exists a unique $v = \Upsilon(u) \in [-K_v, K_v]$ such that $g(u, v) = 0$, and $\Upsilon(u)$ is asymptotically stable for the equation*

$$\frac{dV}{dt} = g(u, V(t)).$$

*Then $u_\epsilon$ converges to a limit $u$ which satisfies almost-everywhere:*

$$\frac{du}{dt} = f(u, \Upsilon(u)).$$

A more detailed result and a proof are given in Chapter 8, Theorem 8.1.

### 4.2.3 Numerical analysis

The numerical integration of slow-fast dynamics raises some well-known issues. Any suitable integration scheme should be asymptotic preserving, which means that one should not need to reduce the time-step overly to maintain convergence as $\epsilon$ is decreased. Many implicit Runge-Kutta schemes satisfy this property (see [98]), which explicit schemes do not. For relevant numerical results, see Chapter 10.

Non-standard finite differences (NSFD) schemes were introduced by Mickens "during the period 1982-1992" (see [11], [168]) to avoid some numerical instabilities, and get numerical solutions whose transient behavior is consistent with what can be known of analytical solutions. Following [169], we have usually chosen a scheme satisfying the Discrete Consistency (DC) principle, *i.e.* producing discrete numerical solutions sharing some qualitative properties with the continuous solutions, in the context of bistable monotone systems (as presented in Section 4.2.1). Namely, we expect the discrete solution to be monotone increasing for all $k \geq k_0$ if it is increasing at step $k_0$, and also to satisfy a comparison principle at the discrete level. A discussion can be found in Chapter 9.

## 4.3 Reaction-diffusion equations

The use of reaction-diffusion (semilinear parabolic) equations in mathematical biology has a long history (see [187, Chapter 1] and the references therein). Reaction-diffusion equations have been introduced simultaneously and independently by Fisher [87] and Kolmogorov, Petrovskii and Piskunov [138] in 1937 in population dynamics modeling. The paper of Fisher is overtly motivated by the propagation of advantageous genes as waves, and seeks "the simplest possible conditions"[2].

Advection-reaction-diffusion equations (or reaction-diffusion equations for short) take the general form of semi-linear parabolic equations, that is:

$$\underbrace{\partial_t \mathbf{n} - \nabla \cdot (A \nabla \mathbf{n})}_{\text{"diffusion": random movement}} + \underbrace{B \nabla \mathbf{n}}_{\text{"advection": drift}} = \underbrace{\mathbf{f}(t, x, \mathbf{n})}_{\text{"reaction": birth and death}},$$

[2] Fisher already notes in the introduction of [87] "The use of the analogy of physical diffusion will only be satisfactory when the distances of dispersion in a single generation are small compared with the length of the wave. In reality, diffusion is a complex process, compounded often of the diffusion of gametes, and that of larvae, in addition to adult forms; a more exact treatment than that supplied by a simple coefficient would involve the interaction of these components, and the stages at which the selective advantage was enjoyed. So far as it is applicable, the analogy of physical diffusion, therefore, greatly simplifies the problem".

combining the random movement through a heat operator, the preferential drift through a $B\nabla$ operator (left-hand side) and birth and death processes through a non-conservative (*i.e.* the total population may vary in time) nonlinearity (right-hand side). The reaction terms can be interpreted in population dynamics as a growth rate, which may depend on the population size $\mathbf{n}$, but also on the location $x$ and of the time $t$ (for instance, of temperature or season).

To fix mathematical terms, let $\Omega \subset \mathbb{R}^d$ be a smooth domain. The reaction-diffusion equations considered here take the form of "Lotka-Volterra systems" (in the sense of [187]):

$$\begin{cases} \partial_t \mathbf{n}(t,x) - \nabla \cdot \big(\mathbf{A}(t,x) \otimes \nabla \mathbf{n}(t,x)\big) + \mathbf{B}(t,x) \otimes \nabla \mathbf{n}(t,x) = \mathbf{n}(t,x) \odot \mathbf{F}(t,x,\mathbf{n}(t,x)) \\ \text{for } (t,x) \in \mathbb{R}_+ \times \Omega, \\ \mathbf{n}(0,x) = \mathbf{n}^0(x) \text{ for } x \in \Omega. \end{cases} \tag{4.7}$$

where $\mathbf{A}, \mathbf{B} : \mathbb{R}_+ \times \Omega \to M_d(\mathbb{R})^{N_d}$, $\mathbf{F} : \mathbb{R}_+ \times \Omega \times \mathbb{R} \to \mathbb{R}$, and the solution is $\mathbf{n} : \mathbb{R}_+ \times \Omega \to \mathbb{R}^{N_d}$.

The basic assumption is that $\mathbf{A}$ is elliptic, that is:

$$\exists \nu > 0,\ \forall i \in [\![1, N_d]\!], \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R}^d,\ \forall \xi \in \mathbb{R}^d,\ \xi^* \mathbf{A}_i \xi \geq \nu |\xi|^2, \tag{Ell}$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$. The assumption (Ell) is obviously satisfied, with $\nu = 1$, if for all $i \in [\![1, N_d]\!]$, $\mathbf{A}_i \equiv \mathbf{I}_d$.

Assuming (Ell) and that the reaction terms in (4.7) are bounded, a non-negativity property holds ([187, Lemma 1.1])):

**Lemma** (Non-negativity principle)**.** *Assume that initial data $\mathbf{n}_i^0 \in L^2(\Omega)$ are nonnegative (for $i \in [\![1, N_d]\!]$), that $\mathbf{B} \equiv 0$ and that there is a locally bounded function $\Gamma(t)$ such that $|\mathbf{F}(t,x,\mathbf{n}(t,x))| \leq \Gamma(t)$ for almost every $x \in \Omega$ along any weak solution $\mathbf{n}$ in $\mathcal{C}^0\big(\mathbb{R}_+, (L^2(\Omega))^{N_d}\big)$. Then these weak solutions satisfy $\mathbf{n}_i \geq 0$.*

Apart from this important qualitative feature (which justifies the use of such equations for population dynamics modeling), not much can be said unless further assumptions are made on the various parameters and functions.

### 4.3.1 Modeling framework

Reaction-diffusion systems of the form (4.7) are considered in Part II. They have at most two components: $N_d \in \{1, 2\}$, representing individuals from the same (mosquito) species that are grouped together depending on their infection status with respect to some *Wolbachia* bacterium (see Chapter 3, in particular Section 3.3.2 for further details).

We also assume that these two groups (which we often call "populations", following the standard use in applied mathematics although they cannot be distinguished as different populations in the biological sense of non-mixing groups) are in competitive interaction, that is:

$$\frac{\partial F_1}{\partial n_2}, \frac{\partial F_2}{\partial n_1} \leq 0.$$

Using continuous densities when dealing with finite populations can only be justified if the population sizes are large enough, and the diffusion approximation can be valid only within large enough time and space scales. These two restrictions must be kept in mind when ecological consequences are drawn from model analysis. Several mitigations could be introduced in order to go beyond these limitations, and in particular the elliptic operator $\nabla \cdot (\mathbf{A}\nabla)$ could be replaced by a more adapted dispersion operator, based on experimental data. This is far beyond the scope of this thesis.

However, the bistable nature of the systems we consider escapes at least one known issue of reaction-diffusion waves (or fronts), that is solutions which propagate a steady state (as the advantageous gene of [87]). The behavior of the so-called "monostable" or "pulled" fronts is determined by the linearization at the edge of the front, precisely where the population is small and the diffusion approximation is not well justified. On the contrary, the "bistable" or "pushed" front that occur in the models studied here are not ruled by their edge linearization, but rather by their profile on a wider area, in a nonlinear fashion. We refer to [92] for a very interesting discussion on the nature of the front (pushed or pulled) and its genetic consequences.

### 4.3.2 Some properties of the scalar equation

We now recall some well-known properties of homogeneous semi-linear scalar reaction-diffusion equations. For more involved dynamical properties and a monotone systems viewpoint, we refer the reader to the review by Polacik [189]. We simplify (4.7) as

$$\partial_t u - \mathcal{L}u = f(u) \text{ in } \Omega, \quad \forall t > 0, u(t, \cdot) = g(\cdot) \text{ on } \partial\Omega, \quad u(0, \cdot) = u^0(\cdot) \text{ in } \Omega, \tag{4.8}$$

where $\mathcal{L}$ is an elliptic operator $\mathcal{L} := \Delta + k(x)\nabla$ and $k, g$ are smooth function $\Omega \to \mathbb{R}$ and $\partial\Omega \to \mathbb{R}$, respectively.

We assume that

$$f \text{ is Lipschitz, } f(0) = 0 \text{ and } f(1) = 0. \tag{4.9}$$

First, we define the sub- and super-solutions in this context (see [192] for a detailed account):

**Definition 4.4.** *A subsolution (resp. a supersolution) of the elliptic problem*

$$-\mathcal{L}u = f(u) \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega \tag{4.10}$$

*is $\underline{u} \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$ (resp. $\overline{u} \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$) such that*

$$-\mathcal{L}\underline{u} \le f(\underline{u}) \text{ in } \Omega, \quad \underline{u} \le g \text{ on } \partial\Omega$$

*(respectively such that*

$$-\mathcal{L}\overline{u} \ge f(\overline{u}) \text{ in } \Omega, \quad \overline{u} \ge g \text{ on } \partial\Omega.)$$

*Similarly, a subsolution (resp. a supersolution) to the parabolic problem (4.8) is any $\underline{u} \in \mathcal{C}^1\big(\mathbb{R}_+; \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})\big)$ such that*

$$\partial_t \underline{u} - \mathcal{L}\underline{u} \le f(\underline{u}) \text{ in } \Omega, \quad \forall t > 0, \underline{u}(t, \cdot) \le g(t, \cdot) \text{ on } \partial\Omega, \quad \underline{u}(0, \cdot) \le u^0(\cdot) \text{ in } \Omega.$$

*(respectively $\overline{u} \in \big(\mathbb{R}_+; \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})\big)$ such that*

$$\partial_t \overline{u} - \mathcal{L}\overline{u} \ge f(\overline{u}) \text{ in } \Omega, \quad \forall t > 0, \overline{u}(t, \cdot) \ge g(t, \cdot) \text{ on } \partial\Omega, \quad \overline{u}(0, \cdot) \ge u^0(\cdot) \text{ in } \Omega.)$$

*By definition, a solution is any function which is simultaneously a sub- and a super-solution.*

Sub- and super-solution provide an iterative scheme to build solutions:

**Proposition 4.5** (Sub- and super-solution method)**.** *Let $\underline{u}$ be a subsolution (respectively $\overline{u}$ a supersolution) to (4.10). If $\underline{u} < \overline{u}$ (which means $\underline{u}(x) \le \overline{u}(x)$ and $\overline{u} \ne \underline{u}$) then there exist minimal and maximal solutions $u_* \le u^*$ such that $\underline{u} \le u_* \le u^* \le \overline{u}$.*

These objects also yields the parabolic comparison principle :

**Proposition 4.6** (Parabolic comparison principle)**.** *For all $T > 0$ we introduce the "parabolic boundary"*

$$\partial_T\Omega := \Big([0, T) \times \partial\Omega\Big) \bigcup \Big(\{0\} \times \Omega\Big).$$

*If $\underline{u}$ (resp. $\overline{u}$) is a sub-solution (resp. a super-solution) to (4.8), and $u$ is a solution such that $u \ge \overline{u}$ (resp. $u \le \underline{u}$) on $\partial_T\Omega$ then the inequality holds on $\Omega \times [0, T]$.*

In addition, the maximum (resp. the minimum) of two sub-solutions (resp. super-solutions) is again a sub-solution (resp. a super-solution). We also define the stability from below and above:

**Definition 4.5.** *A solution $u$ to an elliptic problem is said to be stable from below (resp. above) if for all $\epsilon > 0$ small enough, there exists a subsolution $\underline{u}$ (resp. a supersolution $\overline{u}$) to the problem such that $u - \epsilon \le \underline{u} \le u$. (resp. $u \le \overline{u} \le u + \epsilon$).*

*It is said unstable from below (resp. above) if for all $\epsilon > 0$ small enough there exists a super-solution $\overline{u}$ (resp. a subsolution $\underline{u}$) to the problem such that $u - \epsilon \le \overline{u} \le u$ (resp. $u \le \underline{u} \le u + \epsilon$).*

Some specific nonlinearities have received considerable attention in the mathematical literature. We focus here on two of them[3]:

[3]The ignition nonlinearity ought to be mentioned also. It is defined by a reaction $f$ equal to 0 on $[0, \theta_0]$ and positive on $(\theta_0, 1)$. Among monostable nonlinearities we can highlight the so-called "Fisher-KPP" (after [87] and [138]). They are such that $f$ is $\mathcal{C}^1$ on $[0, \delta]$ for some $\delta > 0$ and for all $p \in (0, 1)$, $0 < f(p) < f'(0)p$. For the Fisher-KPP reaction, there exists a (unique up to translation) traveling wave with speed $c$ for all $c \ge c^* := 2\sqrt{f'(0)}$.

**Definition 4.6.** *We call $f$ **monostable** if, in addition to (4.9), $f > 0$ on $(0,1)$. We call $f$ **bistable** if, in addition to (4.9), there exists $\theta \in (0,1)$ such that $f(\theta) = 0$, $f < 0$ on $(0,\theta)$ and $f > 0$ on $(\theta, 1)$.*

In all cases, we also assume that $f < 0$ on $(-\infty, 0) \cup (1, +\infty)$ (this is a technical assumption to facilitate some proofs, $p$ remains between 0 and 1 if $0 \leq p^0 \leq 1$).

In the bistable case, we also assume without loss of generality (up to changing $p$ into $1-p$) that $\int_0^1 f(x)dx \geq 0$ and define $\theta_c$ as the unique real number in $(0,1]$ such that

$$\int_0^{\theta_c} f(x)dx = 0. \tag{4.11}$$

(Obviously, $\theta_c > \theta$). We define $F(x) = \int_0^x f(\xi)d\xi$, so that $F(\theta_c) = 0$.

When considering compactly supported (or at least localized) initial data for the bistable scalar equation, a sharp threshold principle applies. For instance, with $\Omega = \mathbb{R}$, [70, Theorem 1.3] reads:

**Theorem 4.1.** *Let $\phi_\lambda$, $\lambda > 0$ be a family of $L^\infty(\mathbb{R})$ nonnegative, compactly supported initial data such that*
*(i) $\lambda \mapsto \phi_\lambda$ is continuous from $\mathbb{R}^+$ to $L^1(\mathbb{R})$;*
*(ii) if $0 < \lambda_1 < \lambda_2$ then $\phi_{\lambda_1} \leq \phi_{\lambda_2}$ and $\phi_{\lambda_1} \neq \phi_{\lambda_2}$;*
*(iii) $\lim_{\lambda \to 0} \phi_\lambda(x) = 0$ a.e. in $\mathbb{R}$.*

*Let $p_\lambda$ be the solution to (4.8) with initial data $p_\lambda(0, \cdot) = \phi_\lambda$, and assume that the nonlinearity $f$ is bistable. Then, one of the following alternative holds:*

*(a) $\lim_{t\to\infty} p_\lambda(t,x) = 0$ uniformly in $\mathbb{R}$ for every $\lambda > 0$;*

*(b) there exists $\lambda^* \geq 0$ and $x_0 \in \mathbb{R}$ such that*

$$\lim_{t\to\infty} p_\lambda(t,x) = \begin{cases} 0 & \text{uniformly in } \mathbb{R} \quad (0 \leq \lambda < \lambda^*), \\ u_{\theta_c}(x - x_0) & \text{uniformly in } \mathbb{R} \quad (\lambda = \lambda^*), \\ 1 & \text{locally uniformly in } \mathbb{R} \quad (\lambda > \lambda^*), \end{cases}$$

*where $u_{\theta_c}$ is the unique ground state*[4].

Complementary results on sharp thresholds can also be found in [174] and [163]. The first author to prove sharpness was Zlatos in [243], for the particular case of indicator functions of increasing intervals.

This nice property justifies the interest of studying initial data in the context of localized releases (see Chapter 7), as they can be classified (except for an essentially zero-measure set) into the two groups of "initiating propagation" or "decaying to the original state".

Finally, we discuss the traveling wave solutions, which are wave solutions traveling at a constant speed with a constant shape. They are most simply described in the specific case of the real line (and without drift), when $\Omega = \mathbb{R}$, that is for:

$$\partial_t p - D\partial_{xx} p = f(p) \text{ in } \mathbb{R}_+ \times \mathbb{R}. \tag{4.12}$$

**Definition 4.7.** *A **traveling wave solution** to (4.12) is $(\phi, c)$ where $\phi$ is a **profile** (in $\mathcal{C}^2(\mathbb{R}, [0,1])$) and $c$ is a **speed** (in $\mathbb{R}$) such that $\phi(-\infty) = 1$, $\phi(+\infty) = 0$ and $(t,x) \mapsto \phi(x - ct)$ is an entire solution to (4.12).*

We recall the following fact (see classical literature [85] and [13] or [52] for a more recent proof)

**Proposition 4.7** (Bistable traveling wave)**.** *If $f$ is bistable, then there exists a unique $c = c_*(f)$, and a unique (up to translations) $p_*$ solution of*

$$-p_*'' - cp_*' = f(p_*) \text{ in } \mathbb{R}, \quad p_*(-\infty) = 1,\ p_*(+\infty) = 0.$$

*In addition, $p_*$ is positive and decreasing. We call $c_*$ the bistable wave speed, and $p_*$ the bistable traveling wave, because $u(t,x) = p_*(x - ct)$ is a solution to (6.1) on $\mathbb{R}$.*

The extension of this concept to competitive systems is explained below.

[4] A **ground state** in this context is a solution to $-\partial_{xx} u = f(u)$ such that $u > 0$. See section 7.3.1 below for further details.

### 4.3.3 Some properties of the competitive two-dimensional systems

We now assume that $\Omega = \mathbb{R}^d$ and $N_d = 2$ and consider with homogeneous diffusion $D_1, D_2 > 0$ (which is a specific case of (4.7)) as

$$\begin{cases} \partial_t n_1 - D_1 \Delta n_1 = f_1(n_1, n_2) \text{ in } \mathbb{R}_+ \times \Omega, \\ \partial_t n_2 - D_2 \Delta n_2 = f_2(n_1, n_2) \text{ in } \mathbb{R}_+ \times \Omega, \\ n_i(0, x) = n_i^0(x),\ i \in \{1, 2\}. \end{cases} \tag{4.13}$$

We assume that the solutions to (4.13) are nonnegative and uniformly bounded by $K > 0$ in supremum norm if initial data are nonnegative and bounded by $K$. (This holds typically for Lotka-Volterra reaction terms, as in the example (4.15) below). We consider only competitive systems, that is:

$$\forall n_1, n_2 \in \mathbb{R}_+^2, \quad \frac{\partial f_1}{\partial n_2}, \frac{\partial f_2}{\partial n_1} \leq 0. \tag{4.14}$$

Then, a comparison principle holds (or "operator monotonicity", [187, Lemma 1.2]):

**Lemma** (Comparison principle)**.** *Let $n_{i,j}^0 \in L^2(\Omega)$ such that $0 \leq n_{i,j}^0 \leq K$ for $i, j \in \{1, 2\}^2$, and assume that $f_1, f_2$ along with their first-order partial derivatives are bounded on $[0, K]^2$. Then*

$$n_{1,1}^0 \leq n_{1,2}^0 \text{ and } n_{2,1}^0 \geq n_{2,2}^0 \implies \forall t \geq 0,\ n_{1,1}(t) \leq n_{1,2}(t) \text{ and } n_{2,1}(t) \geq n_{2,2}(t).$$

This lemma makes possible use of a comparison principle as for scalar equations (extending Proposition 4.6), but is limited to monotone reaction terms. In particular, it can be used to study the properties of traveling waves. Restricting to dimension $d = 1$ and to the case of homogeneous diffusion, we consider the problem (1.1) to illustrate this fact:

$$\begin{cases} \partial_t n_1 - D_1 \partial_{xx} n_1 = b_1(1 - s_h \frac{n_2}{n_1+n_2})(1 - \frac{n_1+n_2}{K_1}) - d_1 n_1 \text{ in } \mathbb{R}_+ \times \mathbb{R}, \\ \partial_t n_2 - D_2 \partial_{xx} n_2 = b_2 n_2 (1 - \frac{n_1+n_2}{K_2}) - d_2 n_2 \text{ in } \mathbb{R}_+ \times \mathbb{R}, \\ n_i(0, x) = n_i^0(x) \geq 0,\ i \in \{1, 2\}. \end{cases} \tag{4.15}$$

This system is of Lotka-Volterra form, so that the non-negativity principle applies: for all $t \geq 0$ and $i \in \{1, 2\}$, $n_i(t, \cdot) \geq 0$. Moreover, the reaction terms are both negative as soon as $n_1 + n_2 \geq \max(K_1, K_2) =: K$. From this it can be shown that $n_1, n_2 \leq K$ (more details are shown in Chapter 5, Lemma 5.3). Let $X_1 := K_1(1 - d_1/b_1)$ and $X_2 := K_2(1 - d_2/b_2)$. There are two exclusion steady states, $\mathbf{E}_1 := (X_1, 0)$ and $\mathbf{E}_2 := (0, X_2)$. We assume that these steady states are positive:

$$b_1 > d_1, \quad b_2 > d_2. \tag{$H_1$}$$

The unique coexistence equilibrium $\mathbf{E}_C := (X_1^C, X_2^C)$ is

$$N^C := X_1^C + X_2^C = K_2\big(1 - \frac{d_2}{b_2}\big) = X_2, \quad p^C := \frac{X_2^C}{X_1^C + X_2^C} = \frac{1}{s_h} \frac{X_1 - N^C}{K_1 - N^C}.$$

Under ($H_1$), $\mathbf{E}_C$ is positive if and only if $\frac{X_1 - X_2}{K_1 - X_2} \in (0, s_h)$. We notice that there always holds $X_1 - X_2 < K_1 - X_2$ so in particular if $K_1 < X_2$ then there is no positive coexistence equilibrium. In general we assume

$$X_1 > X_2. \tag{$H_2$}$$

We also state the assumption

$$\frac{X_1 - X_2}{K_1 - X_2} \in (0, s_h). \tag{$H_3$}$$

**Proposition 4.8.** *If ($H_1$) and ($H_3$) hold then there are exactly four steady states for (4.15) standing in the non-negative quadrant: $\mathbf{0}$, $\mathbf{E}_1$, $\mathbf{E}_2$ and $\mathbf{E}_C$. Under ($H_2$), $\mathbf{E}_1$ and $\mathbf{E}_2$ are stable while $\mathbf{E}_C$ is unstable and the trivial steady state $\mathbf{0}$ is unstable in any direction for the reaction (= space-homogeneous) equations.*

*If only ($H_1$) holds then there are only three steady states: $\mathbf{0}$, $\mathbf{E}_1$ and $\mathbf{E}_2$.*

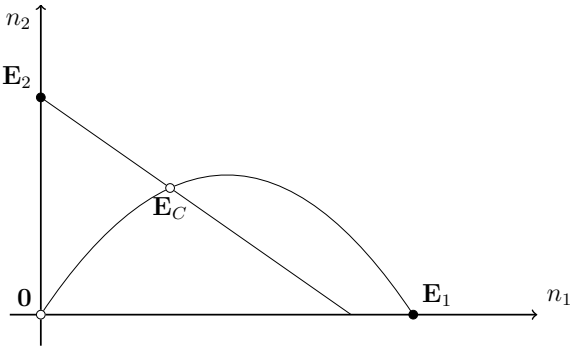



Figure 4.1: Schematic phase diagram for the reaction equations associated with (4.15), showing the nullclines and steady states. Under assumptions ($H_1$), ($H_3$) and ($H_2$), filled (respectively empty) circles stand for stable (respectively unstable) equilibria.

As seen above in the case of scalar equations, traveling-wave solutions on the real line are particular solutions to parabolic equations connecting different steady states at its ends, and moving with constant profile and speed[5]. Here:

**Definition 4.8.** *A traveling-wave solution to (4.15) is a couple $\left(\mathbf{n}^c, c\right)$ where $c \in \mathbb{R}$ and $\mathbf{n}^c = (n_u^c, n_i^c)$ is a couple of monotone functions $\mathbb{R} \to \mathbb{R}_+$, with $\lim_{+\infty} \mathbf{n}^c = \mathbf{E}_1$, $\lim_{-\infty} \mathbf{n}^c = \mathbf{E}_2$ and $(t,x) \mapsto \left(n_u^c(x-ct), n_i^c(x-ct)\right)$ is a solution to (4.15).*

For the bistable system (4.15) we can extend the scalar Proposition 4.7:

**Theorem.** *Under assumptions ($H_1$), ($H_2$), ($H_3$), there exists a unique (up to translation) traveling wave solution $(\mathbf{n}^{c_*}, c_*)$ to (4.15), which is asymptotically stable.*

However, to our best knowledge the sign of $c_*$ can no longer be determined in a simple way (in particular, the issue of nonlinear determinacy - associated with pulled fronts - has been studied in details since [131], see for instance [147] and [96]), but we refer to Chapter 13 for additional discussion.

Pioneering works on this topic include the results of Gardner [91], which cannot be applied directly here[6], and Conley and Gardner [58].

The approach followed in the latter article is suitable here, since the zero sets of $f_1$ and $f_2$ are already in the form [58, Figure 8], satisfy $\int_0^{X_1} f_1(s,0)ds > 0$ and $\int_0^{X_2} f_2(0,s)ds > 0$ and the system

[5] In this light, a traveling-wave solution can also be seen as a heteroclinic orbit of a first-order ODE system of four equations with parameter $c$.

[6] The specific shape assumptions on the functions $M, N$ defined by

$$\begin{cases} M(x,y) := \frac{f_1(x,y)}{x} = b_1(1 - s_h \frac{y}{x+y})(1 - \frac{x+y}{K_1}) - d_1, \\ N(x,y) := \frac{f_2(x,y)}{y} = b_2(1 - \frac{x+y}{K_2}) - d_2, \end{cases}$$

is not satisfied here. Specifically, to apply [91, Theorem 1.2],

- conditions $(i)$ and $(ii)$ of the cited article are met. Indeed, $M, N < 0$ if $\max(x,y) \geq \max(K_1, K_2)$ and

$$\partial_y M = -\frac{b_1}{K_1}(1 - s_h \frac{y}{x+y}) - b_1 s_h \frac{x}{(x+y)^2}(1 - \frac{x+y}{K_1}), \quad \partial_x N = -\frac{b_2}{K_2} < 0,$$

  so $\partial_y M < 0$ in the interesting region of the nonnegative quadrant;
- assumption [91, Theorem 1.2, (b)] is met thanks to Proposition 4.8;
- assumption [91, Theorem 1.2, (c)], which (up to a misprint of the cited article) should read in the vocabulary of the present paper

$$\partial_x f_2(\mathbf{E}_1)^2 < 4\partial_x f_1(\mathbf{E}_1)\partial_y f_2(\mathbf{E}_1), \quad \partial_y f_1(\mathbf{E}_2)^2 < 4\partial_x f_1(\mathbf{E}_2)\partial_y f_2(\mathbf{E}_2),$$

  is satisfied.

But assumption [91, Theorem 1.2, (a)] does not hold. Indeed, the set $\{M = 0\} \cup \mathbb{R}_+^2$ is not the graph of a monotone decreasing function $y = k(x)$: it is rather the graph of an increasing-decreasing function. However, the remainder of the assumption holds: the zero set of $N$ is the graph of the monotone decreasing function $x = l(y) := K_2(1 - \frac{d_2}{b_2}) - y$, and the zero sets of $M$ and $N$ intersect exactly once, at $\mathbf{E}_C$, in the positive quadrant, under assumptions ($H_1$), ($H_2$) and ($H_3$). This suggests that the proof in [91] should be adapted to suit the present problem: Gardner states explicitly [91, p.346] that assumption $(c)$ and the monotonicity of $k$ and $l$ can be relaxed.

is competitive with resource limitation at level $K$. Therefore [58, Theorem p.6] applies and there exists a traveling wave solution to (4.15).

We also have stability and uniqueness of this traveling wave solution. Up to an affine change of variables (replacing $n_2$ by $\widetilde{n}_2 := X_2 - n_2$), we can apply [235, Theorem 2.7] to get uniqueness and [235, Theorem 2.6] to get the global asymptotic stability of this unique traveling wave[7].

In Chapter 13, we state a conjecture (Conjecture 13.1) extending the sharp threshold principle (Theorem 4.1) to such bistable systems, relying on the comparison principle.

### 4.3.4 Numerical analysis

Some numerical simulations or reaction-diffusion systems of the form (4.7) have been performed to illustrate various results throughout Part II, in dimension $d \in \{1, 2\}$ and with $N_d \in \{1, 2\}$. As it gave satisfactory results for our purposes, we have sticked to centered finite-differences scheme for diffusion with Euler implicit time integration.

In short, for (4.12), we discretize the space interval $[-L, L]$ into $N_x + 1$ points $x_i = -L + i\Delta x$ with $i \in [\![0, N_x]\!]$ and $\Delta x = 2L/N_x$. Then, the time interval is discretized into $N_t+1$ times $t_k = k\Delta t$, with $\Delta t = T/N_t$. The simplest scheme for the discrete solution $(n^{k,i})_{\substack{0\le k\le N_t\\ 0\le i\le N_x}}$ reads

$$\begin{cases} \dfrac{n^{k+1,i} - n^{k,i}}{\Delta t} - D\dfrac{n^{k+1,i+1} + n^{k+1,i-1} - 2n^{k+1,i}}{(\Delta x)^2} = f(n^{k,i}),\ k \in [\![0, N_t - 1]\!],\ i \in [\![1, N_x - 1]\!], \\ n^{k,0} = n^{k,1}, \quad n^{k,N_x} = n^{k,N_x-1}, \\ \\ n^{0,i} = n^0(x_i). \end{cases}$$

Introducing $\mathbf{n}^k := (n^{k,i})_{0\le i\le N_x}$, this scheme is written in condensed form as a linear system $\mathbf{M}^1\mathbf{n}^{k+1} = \mathbf{n}^k + \Delta t \cdot \mathbf{f}^k$, where $\mathbf{f}_i^k = f(\mathbf{n}_i^k)$ and with $\xi := \frac{D\Delta t}{(\Delta x)^2}$,

$$\mathbf{M}^1 = \mathbf{M}^1(\xi) := \begin{pmatrix} 1+\xi & -\xi & 0 & \cdots & 0 \\ -\xi & 1+2\xi & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+2\xi & -\xi \\ 0 & \cdots & 0 & -\xi & 1+\xi \end{pmatrix}.$$

This scheme is known to be unconditionally stable (it is studied in [193, Section 12.2]), and $\mathbf{M}^1$ is a symmetric positive-definite $M$-matrix (following[8] [193, Exercise 12.8.2, p. 578]), so that this scheme satisfies a discrete non-negativity principle: if $\mathbf{n}^k + \Delta t \cdot \mathbf{f}^k \ge 0$ then $\mathbf{n}^{k+1} \ge 0$. Therefore, in the bistable case a sufficient condition for positivity of $(n^{k,i})_{k,i}$ is that

$$\Delta t \le \inf_{p\in(0,\theta)} \frac{p}{-f(p)}. \tag{4.16}$$

Likewise, under a condition on $\Delta t$ the comparison principle holds at the discrete level since we get $\mathbf{M}^1(\mathbf{n}_1^k - \mathbf{n}_2^k) = \mathbf{n}_1^k - \mathbf{n}_2^k + \Delta t \cdot (\mathbf{f}_1^k - \mathbf{f}_2^k)$, so that $\mathbf{n}_1^0 \ge \mathbf{n}_2^0$ implies $\mathbf{n}_1^k \ge \mathbf{n}_2^k$ for all $k \in \mathbb{Z}_{\ge 0}$ if

$$\Delta t \le \frac{-1}{\inf_{p\in[0,1]} \min(f'(p), 0)}. \tag{4.17}$$

(Note that assuming that $f$ is of class $\mathcal{C}^1$ with $f(0) = 0 = f(1)$ implies that the right-hand side is positive.)

In addition, the constant vector $\mathbb{1}$ is an eigenvector of $\mathbf{M}$ associated with the eigenvalue 1, so that constants are fixed points for this scheme. Combining this with the discrete comparison principle shows that the solutions remain upper bounded by $K > 0$ if the initial data is upper bounded by $K$ such that $f(K) = 0$.

This analysis can be extended to systems, where we typically write the scheme as

$$\mathbf{M}_1^1\mathbf{n}_1^{k+1} = \mathbf{n}_1^k + \Delta t \cdot \mathbf{f}_1^k, \quad \mathbf{M}_2^1\mathbf{n}_2^{k+1} = \mathbf{n}_2^k + \Delta t \cdot \mathbf{f}_2^k,$$

[7] It seems that these results could be proved by applying the methods in [52] to monotone systems.

[8] In fact, in our case the matrix is slightly different but the proof is the same, it suffices to show that $\mathbf{M}^1 x + \alpha x \ge 0 \implies x \ge 0$ for $\alpha > 0$, and then to pass to the limit $\alpha \to 0$.

with $[\mathbf{f}_j^k]_i = f_j([n_1^k]_i, [n_2^k]_i)$ for $j \in \{1,2\}$ and $i \in [\![1, N_x]\!]$, and where $\mathbf{M}_i^1 = \mathbf{M}^1(\frac{D_i \Delta t}{(\Delta x)^2})$. Condition (4.16) rewrites in this case

$$\forall i \in \{1,2\}, \quad \Delta t \leq \inf_{(n_1,n_2)\in[0,K]^2} \frac{-n_i}{\min(f_i(n_1,n_2),0)},$$

under which assumption the non-negativity property holds for the system. The discrete comparison principle can also be stated in this context under the assumption (extending (4.17)):

$$\forall i \in \{1,2\}, \Delta t \leq \frac{-1}{\inf_{(n_1,n_2)\in[0,K]^2} \min(\partial_i f_i(n_1,n_2),0)}.$$

As in the scalar case, by applying the discrete comparison principle we have that if the initial data belongs to some interval $[\mathbf{E}_2, \mathbf{E}_1]$ for the order induced by $\mathcal{K}^o = \mathbb{R}_+^{N_x+1} \times \mathbb{R}_-^{N_x+1}$, with $\mathbf{E}_1 \geq_{\mathcal{K}^o} \mathbf{E}_2$, then so does the discrete solution.

We emphasize that the numerical resolution of linear systems $\mathbf{M}^1 X = Y$ does not require the inversion of $\mathbf{M}^1$ and can be done in $O(N_x)$ computations using the Thomas algorithm (see [193, Section 3.7.1]).

The two-dimensional analogue of this scheme is built on a rectangular grid, thanks to the pentadiagonal matrix

$$\mathbf{M}^2 = \mathbf{M}^2(\xi) := \begin{pmatrix} M_{2,3}^{N_x}(\xi) & \xi \mathbf{I}_{N_x+1} & 0 & \cdots & 0 \\ \xi \mathbf{I}_{N_x+1} & M_{3,4}^{N_x} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & M_{3,4}^{N_x} & \xi \mathbf{I}_{N_x+1} \\ 0 & \cdots & 0 & \xi \mathbf{I}_{N_x+1} & M_{2,3}^{N_x} \end{pmatrix} \in M_{(N_x+1)\cdot(N_y+1)}(\mathbb{R}),$$

where we define

$$M_{\alpha,\beta}^{N_x}(\xi) := \begin{pmatrix} 1+\alpha\xi & -\xi & 0 & \cdots & 0 \\ -\xi & 1+\beta\xi & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+\beta\xi & -\xi \\ 0 & \cdots & 0 & -\xi & 1+\alpha\xi \end{pmatrix} \in M_{N_x+1}(\mathbb{R}).$$

Again, we note that $\mathbb{1}_{(N_x+1)\cdot(N_y+1)}$ is an eigenvector of $\mathbf{M}^2$ associated with the eigenvalue 1, so the constants are fixed points of the scheme. The previous properties are extended to this case upon noting that $\mathbf{M}^2$ is again a symmetric positive-definite $M$-matrix.

To check this fact, we simply check that for $x \in \mathbb{R}^{(N_x+1)\cdot(N_y+1)}$,

$$x^* \mathbf{M}^2 x = \sum_{i=1}^{N_x} \sum_{j=0}^{N_y-1} x_{i+jN_x}^2 + \xi\Big( \sum_{i=1}^{N_x} \sum_{j=0}^{N_y-1} (x_{i+jN_x} - x_{i-1+jN_x})^2 + \sum_{i=0}^{N_x} \sum_{j=1}^{N_y-1} (x_{i+jN_x} - x_{i+(j-1)N_x})^2 \Big),$$

so that $\mathbf{M}$ is positive definite. It is also diagonally dominant. Finally, the trick from [193, Exercise 12.8.2, p. 578] also applies here, yielding the conclusion.

Here also, a generalization of Thomas algorithm allows the resolution of the pentadiagonal linear system in an efficient way (that is, in $O(N_x \cdot N_y)$ operations).

## 4.4 Important auxiliary results

For the completeness of the context exposition, we gather here some classical mathematical results that are used during the thesis, although they do not necessarily come from population dynamics, originally.

### 4.4.1 Perron-Frobenius theory

In various places and in particular in Chapter 11, Metzler matrices arise as natural objects.

**Definition 4.9.** *A **Metzler matrix** is a square matrix $A \in M_n(\mathbb{R})$ whose off-diagonal coefficients are non-negative, $A_{ij} \geq 0$ for all $i \neq j$.*

A matrix $A \in M_n(\mathbb{R})$ is called irreducible if it is not similar by a permutation to a block upper triangular matrix. Defining the graph $G$ with vertices $(V_i)_{i\in[\![1,n]\!]}$ and a directed edge from $V_i$ to $V_j$ if and only if $A_{ij} \neq 0$, it is well-known that $A$ is irreducible if and only if $G$ is strongly connected (which means that it contains a directed path from $i$ to $j$ for any $i, j \in [\![1, n]\!]$).

Since some translate $A + \alpha\mathbf{I}$ ($\alpha \in \mathbb{R}$) of a Metzler matrix is a non-negative matrix (*i.e.* $A + \alpha\mathbf{I}$ has non-negative coefficients and is non-zero), the Perron-Frobenius theorem for positive (or non-negative irreducible) matrices extends to Metzler matrices.

We merely state the result and refer to [31] for a proof:

**Theorem 4.2** (Perron-Frobenius)**.** *Let $A \in M_n(\mathbb{R})$ be an irreducible Metzler matrix. Then $\mu(A)$ is an eigenvalue of $A$ associated with a strictly positive eigenvector $V \gg 0$, called the Perron eigenvector of $A$. The eigenvalue $\mu(A)$ is simple and there is no other eigenvector of $A$ which is strictly positive.*

*If $A$ is positive then the same property applies, with $\mu(A) = \rho(A)$.*

Since this theorem also applies to the adjoint $A^*$, it is natural to normalize the strictly positive right and left Perron eigenvectors of $A$ (that is, the Perron eigenvectors of $A$ and $A^*$), respectively $V$ and $V_*$, by $\langle V, V_*\rangle = 1$.

This useful property is extended to infinite-dimensional compact and positive linear operators on Banach spaces by the Krein-Rutman theorem (proved in [139], see the book [69]).

### 4.4.2 Lyapunov functions

We recall some useful definitions and the state and prove an abstract result on Lyapunov functions, which is used in Chapter 12 for a population structured by its phenotype belonging to a metric space $\mathcal{P}$.

For a dynamic $\dot{y} = V(y)$ defined on a metric space $Y$ such that any orbit is relatively compact, we call $f : Y \to \mathbb{R}$ a **global Lyapunov function** for this dynamic if $f$ is continuous, Fréchet-differentiable and $t \mapsto f(y(t))$ is increasing along any orbit of the $V$-dynamic, with strict monotonicity except if $V(y) = 0$ (that is, at **rest points** of $V$).

For any $f : Y \to \mathbb{R}$ we call $A \subset Y$ a **local maximizer set** of $f$ if $f(A)$ is a singleton $\{f_A\}$, $A$ is connected and there exists a neighborhood $B$ of $A$ in $Y$ such that for all $y \in B\backslash A$, $f(y) < f_A$.

We call $f$ a **strict Lyapunov function** for the set $A \subset Y$ if it is a global Lyapunov function and if in addition, $A$ is a local maximizer set of $f$ and there exists a neighborhood $B$ of $A$ such that for all $y \in B\backslash A$, $\langle Df_y, V(y)\rangle > 0$.

If a dynamic is given on $Y$, we say that $A$ is **Lyapunov-stable** (with respect to this dynamic) if every neighborhood $B$ of $A$ contains a neighborhood $B'$ of $A$ such that if $y_0 \in B'$, then for all $t > 0$, $y(t) \in B$. We say that $A$ is **asymptotically stable** if there exists a neighborhood $B$ of $A$ such that if $x \in B$ then $\omega(x) \subseteq A$.

The following is adapted from [198, Theorem 4.4] (the first part is [155, Proposition 1]):

**Proposition 4.9.** *Let $Y$ be a complete metric space on which a dynamic $\dot{y} = V(y)$ is given, such that for all $y_0 \in Y$ the orbit $\{y(t), t \geq 0\} \subset Y$ is relatively compact. Let $J : Y \to \mathbb{R}$ be a global Lyapunov function for this dynamic.*

*Then, the $\omega$-limit set of any $y_0 \in Y$ is non-empty, compact, connected, consists entirely of rest points of $V$ and $f(\omega(y_0))$ is a singleton.*

*In addition, if $A \subset Y$ is a local maximizer set of $J$ then it is Lyapunov-stable.*

*Finally, if $J$ is a strict Lyapunov function for the set $A \subset Y$ then $A$ is asymptotically stable.*

This type of convergence result has appeared in the economic literature devoted to game theory with continuous strategy space, which we denote here by $\mathcal{P}$ (for "phenotype"). For instance it is stated in [55, Theorem 3.a] and follows from [54, Theorem 2]. However, the main part of the proof is to be found in [199] and the gist is already contained in the paper [198, Theorem 4.4] (and the appendix of the cited paper exhibits a proof).

We want to apply Proposition 4.9 for a dynamic defined on $\mathcal{M}_+^1(\mathcal{P})$, equipped with weak* topology. In all the cited papers, the set of strategies $\mathcal{P}$ is assumed to be compact, and the equivalent of Proposition 4.9 is stated under this restriction, which is required in order to get compactness for the weak convergence topology on probability measures. Thanks to a theorem of Prohorov, we can relax this assumption into $\mathcal{P} \subseteq \mathbb{R}^d$ (precisely, $\mathcal{P}$ is a locally compact Hausdorff space), provided that the dynamic preserves some tightness of the measures: what we need to prevent is the loss of mass at infinity. We define the tight measures set $\widehat{\mathcal{M}}_+^1(\mathcal{P}) \subset \mathcal{M}_+^1(\mathcal{P})$ as:

$$\widehat{\mathcal{M}}_+^1(\mathcal{P}) := \left\{ q_0 \in \mathcal{M}_+^1(\mathcal{P}),\ \forall \epsilon > 0, \exists K \subset \mathcal{P} \text{ compact s.t. } q_0(\mathcal{P}\backslash K) < \epsilon \right\}$$

Then, we assume that $\mathcal{P}$ is a locally compact Hausdorff space and that the $V$-dynamic is uniformly tight, that is:

$$\forall q_0 \in \widehat{\mathcal{M}}_+^1(\mathcal{P}),\ \forall \epsilon > 0, \exists K \subset \mathcal{P} \text{ compact s.t. } \forall t \geq 0, q(t)(\mathcal{P}\backslash K) < \epsilon. \tag{4.18}$$

Under these assumptions, we claim that Proposition 4.9 applies to $Y = \widehat{\mathcal{M}}_+^1(\mathcal{P})$. By Prohorov's theorem and the assumption (4.18), any orbit from an initial data in $\widehat{\mathcal{M}}_+^1(\mathcal{P})$ is relatively compact in the weak* topology when $\mathcal{P}$ is locally compact Hausdorff. This implies that $\omega(q_0)$ is non-empty, and the remaining of Proposition 4.9 follows from the same arguments as in the classical case where $\mathcal{P}$ is assumed to be compact.

*Proof of Proposition 4.9.* Let $y_0 \in Y$. Since $t \mapsto f(y(t))$ is an increasing function $\mathbb{R}_+ \to \mathbb{R}$, and is bounded (since it is continuous and the orbit of $y_0$ is relatively compact), it must converge to some $\overline{f} \in \mathbb{R}$.

The set $\omega(y_0)$ is non-empty because the closure of the closure of the orbit $\mathrm{Adh}\{y(t), t \geq 0\}$ is compact, hence it has a cluster point by Bolzano-Weierstrass theorem. Moreover, it is compact and connected because

$$\omega(y_0) = \bigcap_{t \geq 0} \mathrm{Adh}\{y(s),\ s \geq t\},$$

and therefore it writes as the decreasing intersection of connected and compact sets.

Let $y_1 \in \omega(y_0)$. Since $y_1 = \lim_{n \to +\infty} y(t_n)$ for some increasing sequence $(t_n)_n$, by continuity of $f$ we can write

$$\overline{f} = \lim_{t \to +\infty} f(y(t)) = \lim_{n \to +\infty} f(y(t_n)) = f(\lim_{n \to +\infty} y(t_n)) = f(y_1).$$

Hence $f(\omega(y_0))$ is a singleton.

Then, by construction $\omega(y_0)$ is invariant under the $V$-dynamic. Since $f$ is strictly increasing along orbits unless evaluated at rest points of the $V$-dynamic, it implies that $\omega(y_0)$ consists of rest points of the $V$-dynamic, that is $V(\omega(y_0)) = \{0\}$.

Now, let $A \subset Y$ be a local maximizer set of $f$. We want to prove that it is Lyapunov-stable.

Let $B$ be a neighborhood of $A$. By definition of local maximizers, there exists a neighborhood $C \subset B$ of $A$ such that for all $y \in C\backslash A$, $f(y) < f_A$, and $\mathrm{Adh}(C) \subset B$ (using the compactness of $Y$). Let $B'_\epsilon := B \cap \{f_A - \epsilon < f < f_A\}$. By continuity of $f$ and the local maximizer property, for all $\epsilon > 0$, $C \cap B'_\epsilon \neq \emptyset$ and $B'_\epsilon$ is an open subset of $B$. Let $\epsilon > 0$ be small enough so that $\mathrm{Adh}(B'_\epsilon) \subset C$ (namely, $\max_{\mathrm{Adh}(C)\backslash C} f < f_A - \epsilon$), and let $y_0 \in B'_\epsilon$. Then for all $t > 0$ we have $f(y(t)) > f_A - \epsilon$. Let $t_0 := \inf\{t \geq 0, y(t) \notin B'_\epsilon\}$. If $t_0 < +\infty$ then we find that necessarily $f(y(t_0)) = f_A$. Since $y(t_0) \in \mathrm{Adh}(B'_\epsilon) \subset C$, this implies that $y(t_0) \in A$, and so $y(t) \in A \subset B$ for all $t \geq t_0$. Indeed, $A$ is stable under the $V$-dynamic.

All in all, either $t_0 = +\infty$ and $y(t) \in B'_\epsilon \subset B$ for all $t \geq 0$, or $t_0 < +\infty$ and $y(t) \in B'_\epsilon \cup A \subset B$ for all $t \geq 0$, whence Lyapunov-stability of $A$.

For the last point, we simply need to say that if for some $x \in B$ (where the neighborhood $B$ of $A$ is given by the definition of the strict Lyapunov function $f$), there exists $y \in \omega(x) \cap (B\backslash A)$, then $y$ cannot be a rest point of the $V$-dynamic since $\langle Df_y, V(y) \rangle > 0$, which is a contradiction. Thanks to Lyapunov stability, upon shrinking $B$ we know that for all $x \in B$, $\omega(x) \subset B$. Since $\omega(x) \cap (B\backslash A) = \emptyset$ and $\omega(x)$ is non-empty, we can conclude that $\omega(x) \subset A$. □

# Part II

# Reaction-diffusion models

# Chapter 5

# Reduction to a single equation for some 2-by-2 systems

> [...] – comme si l'oeil enchanté, au matin de la création, eût pu voir se dérouler le mystère naïf de la *séparation des éléments.*
>
> Julien Gracq, *Au Château d'Argol.*

This chapter is a joint work with Nicolas Vauchelet. It was published as an article in SIAM Journal on Applied Mathematics [211].

**Abstract.** We consider general models of coupled reaction-diffusion systems for interacting variants of the same species. When the total population becomes large with intensive competition, we prove that the frequencies (*i.e.* proportions) of the variants can be approached by the solution of a simpler reaction-diffusion system, through a singular limit method and a relative compactness argument. As an example of application, we retrieve the classical bistable equation for *Wolbachia*'s spread into an arthropod population from a system modeling interaction between infected and uninfected individuals.

## 5.1 Introduction

We are interested in modeling situations when two biological populations of the same species interact with each other, especially move, reproduce and compete. The dynamics of these two populations are commonly described by a reaction-diffusion system of two equations in the whole space $\mathbb{R}^d$ ($d \geq 1$). In this setting, reaction terms encompass the whole interaction. Usually, they are non-linear, in order to account for competition or mutualistic interaction. Denoting $n_1(t,x)$ and $n_2(t,x)$ the densities of each species' variant at time $t > 0$ and position $x \in \mathbb{R}^d$, the mathematical model reads:

$$\begin{cases} \partial_t n_1 - \nabla \cdot (A(x)\nabla n_1) &= n_1 f_1(n_1, n_2), \\ \partial_t n_2 - \nabla \cdot (A(x)\nabla n_2) &= n_2 f_2(n_1, n_2), \end{cases} \tag{5.1}$$

where the diffusion matrix $A$ is elliptic and the regular functions $f_1$ and $f_2$ describe the interaction between variants. This system is complemented with initial conditions. Since the analysis of such systems is actually delicate, one prefers considering the proportion of one population, for instance $p = \frac{n_1}{n_1+n_2}$. Then the interactions are described through the dynamics of the proportion $p$ by a reaction-diffusion system:

$$\partial_t p - \nabla \cdot (A(x)\nabla p) = pF(p). \tag{5.2}$$

Since the pioneering works of Fisher [87] and Kolmogorov, Petrovskii, Piskunov [138], this kind of reaction-diffusion equation has been extensively studied in mathematical literature. In particular many effort have been done to establish the existence of traveling waves and to describe the invasion phenomena (see e.g. [84], [230]). However, when considering systems of reaction-diffusion equations, many difficulties make such analysis harder. For instance, we mention the work [91] for competitive system. The aim of this paper, is to focus on the link between system (5.1) and (5.2).

More precisely, the main question we want to address is to know if solutions of system (5.1) can be rigorously approximated by system (5.2) for the proportion $p$ of one species. In our main result, we show that under suitable assumptions on the reaction terms in (5.1), the proportion $p = \frac{n_1}{n_1+n_2}$ is close (in a sense which will be defined below) to a solution to system (5.2). More precisely, we show that when the total population becomes large with intensive competition, the frequency $p = \frac{n_1}{n_1+n_2}$ for system (5.1) converges to the solution of equation (5.2) where the non-linear function in the right hand side $F$ is explicitly given with $f_1$ and $f_2$. Our proof is based on a compactness argument resulting from a priori estimates. The closest results of model reduction for competition-diffusion systems, are those of [111] and [112] (in bounded domains, with a specific and extensive discussion on the boundary issues).

Our first interest in this topic comes from the biological phenomenon of cytoplasmic incompatibility, caused by the endo-symbiotic bacterium *Wolbachia* in some arthropod species (see [233], [29], [121]). These bacteria have gained interest lately because of their potential use as a tool to fight arboviruses (see [118], [232]). For this situation, modeled by a reaction-diffusion system, we prove that if reaction terms scale in a proper way, then the frequency of *Wolbachia* infection approaches the solution of a single closed reaction-diffusion equation, which is bistable. Bistable equations have been suggested long ago for this problem (see [29] for an account on this topic, and [203] for a specific discussion). When these models encompass a space-dependent total population density $\rho$ (as proposed e.g. in [177, 26, 29]), they read

$$\partial_t p - \nabla \cdot (A(x)\nabla p) - 2\frac{\nabla \rho}{\rho} A(x)\nabla p = pF(p). \tag{5.3}$$

In some sense our result justifies their use thanks to a rigorous singular limit method. We do not assume that $\rho$ and $p$ vary independently, and find that (5.3) must be corrected since $F$ is a function of $p$ and $\rho$. We warn the reader that in order to simplify the computations, we will define a "reduced total population density" $n$, instead of using the total population density $\rho$ directly.

The outline of the paper is the following. In the next Section, we present the setting of the problem. In particular the assumptions on the reaction terms and the main result are presented. Section 5.3 is devoted to an example of application: the interaction between an infected and an uninfected mosquitoes population. A numerical illustration is also provided in dimension $d = 1$. The proof of our main result is provided in Section 5.4. This proof relies strongly on *a priori* estimates that make us able to prove relative compactness of solutions families when a parameter describing the size of the population goes to $+\infty$. We give in Section 5.5 some extension to our main result. Finally, Section 5.6 highlights questions this work opens.

## 5.2 Setting of the problem for typical Lotka-Volterra systems

In this section, we first define the setting where our result applies (typical Lotka-Volterra systems), and then state it in Theorem 5.1.

### 5.2.1 System and assumptions

For $\epsilon > 0$, let $f_1^\epsilon, f_2^\epsilon : \mathbb{R}^2 \to \mathbb{R}$ be two functions. We start from the following system in $\mathbb{R}^d$

$$\begin{cases} \partial_t n_1^\epsilon - \nabla \cdot (A(x)\nabla n_1^\epsilon) &= n_1^\epsilon f_1^\epsilon(n_1^\epsilon, n_2^\epsilon), \\ \partial_t n_2^\epsilon - \nabla \cdot (A(x)\nabla n_2^\epsilon) &= n_2^\epsilon f_2^\epsilon(n_1^\epsilon, n_2^\epsilon), \end{cases} \tag{5.4}$$

with given initial data $n_i^\epsilon(t = 0, x) = n_i^{\text{init},\epsilon} \geq 0$ for $i \in \{1, 2\}$. We assume that the matrix $A$ is elliptic and that $f_1^\epsilon, f_2^\epsilon$ are smooth enough to guarantee existence and uniqueness of a global solution for fixed $\epsilon > 0$. More precisely,

**Assumption 5.1** (Ellipticity and symmetry of $A$)**.** *The diffusion matrix $A : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is symmetric and the system (5.4) is uniformly elliptic, i.e.*

$$\exists \nu_0 \in \mathbb{R}_+^*, \forall x, \zeta \in \mathbb{R}^d,\ \zeta \cdot (A(x)\zeta) \geq \nu_0 |\zeta|^2,$$

*where $|\cdot|$ stands for the euclidean norm in $\mathbb{R}^d$.*

We define "reduced total population" $n^\epsilon$ and frequency (*i.e.* proportion of population 1) $p^\epsilon$ by

$$n^\epsilon := \frac{1}{\epsilon} - n_1^\epsilon - n_2^\epsilon, \quad p^\epsilon := \frac{n_1^\epsilon}{n_1^\epsilon + n_2^\epsilon}. \tag{5.5}$$

Since 0 is a sub-solution for each equation in (5.4), and since initial data are nonnegative, we have $n_i^\epsilon(t,\cdot) \geq 0$, for any $t \geq 0$. By convention, we take $p^\epsilon = 0$ whenever $n_1^\epsilon = n_2^\epsilon = 0$.

We want to compute the limit as $\epsilon \to 0$ of the frequency $p^\epsilon$ under the above assumption on $A : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and on some assumptions on the families of functions $(f_1^\epsilon, f_2^\epsilon)_{\epsilon>0}$.

As a typical Lotka-Volterra system, we note that absence of either population of type 1 or 2 is a solution to this system: there is no spontaneous generation of one population from the other. In addition, the system is positive: for non-negative initial data, $0 \leq p^\epsilon \leq 1$. Now, we state our key assumptions.

**Assumption 5.2** (Dependence in $\epsilon$)**.** *Functions $f_1^\epsilon, f_2^\epsilon$ are of class $\mathcal{C}^2(\mathbb{R}^2_+ - \{0\})$, and for $i \in \{1,2\}$ there exists $F_i \in \mathcal{C}^2(\mathbb{R}^2_+)$ (independent of $\epsilon > 0$) such that*

$$f_i^\epsilon(n_1^\epsilon, n_2^\epsilon) = F_i(n^\epsilon, p^\epsilon). \tag{5.6}$$

*In other words, for any $n_1, n_2 \geq 0$, we may write $f_i^\epsilon(n_1,n_2) = F_i(\frac{1}{\epsilon} - n_1 - n_2, \frac{n_1}{n_1+n_2})$.*

From now on we drop, the superscript $\epsilon$ when it is not equivocal.

Adding the two equations in system (5.4) and using also the identity

$$\nabla\cdot(A(x)\nabla p) = \frac{1}{n_1+n_2}\nabla\cdot(A(x)\nabla n_1) + \frac{p}{n_1+n_2}\nabla\cdot(A(x)\nabla n) + 2\frac{1}{n_1+n_2}\nabla p\cdot(A(x)\nabla n),$$

we deduce, after straightforward computations, that $(n,p)$ satisfies

$$\begin{cases} \partial_t n - \nabla\cdot(A(x)\nabla n) = \big(n - \frac{1}{\epsilon}\big)\big(pF_1(n,p) + (1-p)F_2(n,p)\big), \\ \partial_t p - \nabla\cdot(A(x)\nabla p) + 2\nabla p\cdot\dfrac{A(x)\nabla n}{\frac{1}{\epsilon} - n} = p(1-p)\big(F_1(n,p) - F_2(n,p)\big), \end{cases} \tag{5.7}$$

complemented with well-defined initial data. According to the first equation in (5.7), it appears interesting, when $\epsilon \to 0$, to consider the function

$$H(n,p) := -pF_1(n,p) - (1-p)F_2(n,p). \tag{5.8}$$

The following assumption guarantees existence of zeros $(n,p) = (h(p),p)$ for each $p \in [0,1]$ for the above function $H$.

**Assumption 5.3** (Nature of the interaction)**.** *In addition to Assumption 5.2, we assume*

*(i)* $\exists B > 0$ *such that* $\forall n \geq 0$, $\forall p \in [0,1]$, $\partial_n H(n,p) \leq -B$ ;

*(ii)* $\forall p \in [0,1]$, $H(0,p) > 0$.

*Conditions (i) and (ii) imply that for all $p \in [0,1]$, there exists a unique $n =: h(p) \in \mathbb{R}^*_+$ such that $H(n,p) = 0$. We assume $H \in \mathcal{C}^2(\mathbb{R}^2_+)$ (which is true if Assumption 5.2 holds), and thus $h \in \mathcal{C}^2(0,1;\mathbb{R})$, with $H(h(p),p) = 0$ for all $p \in [0,1]$.*

In particular, $h(0) = \frac{1}{\epsilon} - \overline{n}_2^\epsilon$, obtained at the population 1-free equilibrium $(0, \overline{n}_2^\epsilon)$ for (5.4). By Assumption 5.3, this equilibrium is unique and the reduced population $n^\epsilon$ does not depend on $\epsilon$.

Assumption 5.3 may seem a little awkward, therefore we would like to point out a sufficient condition.

**Lemma 5.1.** *We assume that both $f_1^\epsilon$ and $f_2^\epsilon$ are smooth (say, of class $\mathcal{C}^2(\mathbb{R}^2_+ - \{(0,0)\})$). We define the "triangle" $T_\epsilon = \{(n_1,n_2) \in \mathbb{R}^2_+$ such that $n_1 + n_2 \leq \frac{1}{\epsilon}\}$. In addition to Assumption 5.2, if $f_1^\epsilon, f_2^\epsilon$ satisfy the following inequalities in $T_\epsilon$,*

$$\forall \underline{n} = (n_1,n_2) \in T_\epsilon, \quad n_1^2\partial_{n_1}f_1^\epsilon(\underline{n}) + n_1n_2(\partial_{n_2}f_1^\epsilon + \partial_{n_1}f_2^\epsilon)(\underline{n}) + n_2^2\partial_{n_2}f_2^\epsilon(\underline{n}) \leq -B(n_1+n_2)^2, \tag{5.9}$$

*together with a "boundary condition": for all $\underline{n} \in \mathbb{R}^2_+$ with $\|\underline{n}\|_1 = \frac{1}{\epsilon}$,*

$$n_1 f_1^\epsilon(\underline{n}) + n_2 f_2^\epsilon(\underline{n}) < 0. \tag{5.10}$$

*Then Assumption 5.3 holds.*

**Remark 5.1.** *Equation (5.9) on the lines $n_1 = 0$ and $n_2 = 0$ means that the profiles of $f_1, f_2$ are below concave parabolic profiles. More generally, it ensures that the total population $n_1 + n_2$, in (5.7), will start decreasing (in time) before reaching the value $\frac{1}{\epsilon}$.*

*Proof of Lemma 5.1.* We verify each point *(i)* and *(ii)* in Assumption 5.3. *(i)* We first recall that $\partial_n H = -(p\partial_n F_1 + (1-p)\partial_n F_2)$. From (5.6), we express $\partial_{n_i} f_1^\epsilon$ and $\partial_{n_i} f_2^\epsilon$,

$$\partial_{n_i} f_j^\epsilon = -\partial_n F_j + \frac{n_{3-i}}{n_1+n_2}\partial_p F_j, \qquad i,j = 1,2.$$

Collecting these expressions yields straightforwardly

$$\begin{aligned}\partial_n H = -p\partial_n F_1 - (1-p)\partial_n F_2 = \quad & p\frac{n_1}{n_1+n_2}\partial_{n_1} f_1^\epsilon + p\frac{n_2}{n_1+n_2}\partial_{n_2} f_1^\epsilon \\ & +(1-p)\frac{n_1}{n_1+n_2}\partial_{n_1} f_2^\epsilon + (1-p)\frac{n_2}{n_1+n_2}\partial_{n_2} f_2^\epsilon,\end{aligned}$$

whence the equivalence with (5.9).

*(ii)* For the boundary condition, we compute from (5.8)

$$H(0,p) = -\big(pF_1(0,p) + (1-p)F_2(0,p)\big).$$

Then it suffices to recall that, by definition in (5.6), for $i \in \{1,2\}$, $F_i(0,p) = f_i^\epsilon(\frac{p}{\epsilon}, \frac{1-p}{\epsilon})$. □

### 5.2.2 Main result

We are now in position to state our main result. We recall that we associate to any initial data $n_i^{\text{init},\epsilon}$ the corresponding solutions of (5.4), $(n_i^\epsilon)$, and their relative variable $n^\epsilon$ and $p^\epsilon$, as defined in (5.5). In addition we may define

$$n^{\text{init},\epsilon} = \frac{1}{\epsilon} - n_1^{\text{init},\epsilon} - n_2^{\text{init},\epsilon}, \quad p^{\text{init},\epsilon} = \frac{n_1^{\text{init},\epsilon}}{n_1^{\text{init},\epsilon} + n_2^{\text{init},\epsilon}}.$$

**Theorem 5.1.** *We assume that Assumptions 5.1, 5.2, and 5.3 are satisfied. We consider the solutions of (5.4) with initial data $n_i^\epsilon(t=0) = n_i^{init,\epsilon} \in L^\infty(\mathbb{R}^d;\mathbb{R}_+)$ for $i \in \{1,2\}$. We assume moreover that there exists $p^{init} \in L^2(\mathbb{R}^d)$ such that*

$$p^{init,\epsilon} \underset{\epsilon\to 0}{\rightharpoonup} p^{init} \text{ in } L^2(\mathbb{R}^d) - \text{weak}, \quad n^{init,\epsilon} - h(0) \in L^2 \cap L^\infty(\mathbb{R}^d), \tag{5.11}$$

*with uniform bounds in $\epsilon > 0$.*

*Then, for all $T > 0$, defining $\mathcal{H}_T^1 = L^2(0,T;L^2(\mathbb{R}^d))$ and $\mathcal{H}_T^2 = L^2(0,T;H^1(\mathbb{R}^d))$, we have the convergence*

$$\begin{cases} p^\epsilon \xrightarrow[\epsilon\to 0]{} p^0 \text{ strongly in } \mathcal{H}_T^1, \text{ weakly in } \mathcal{H}_T^2, \\ n^\epsilon - h(p^0) \xrightarrow[\epsilon\to 0]{} 0 \text{ strongly in } \mathcal{H}_T, \text{ weakly in } \mathcal{H}_T^2, \end{cases} \tag{5.12}$$

*where $p^0$ is the unique solution of the following initial value problem*

$$\begin{cases} \partial_t p^0 - \nabla\cdot(A(x)\nabla p^0) = p^0 F_1(h(p^0), p^0), \\ p^0(t=0) = p^{init}. \end{cases} \tag{5.13}$$

This result asserts that, locally in time, the proportion of the first population, $p$, solution to system (5.4), under suitable assumption on the reaction term and on the initial data, is close to the solution of a single reaction-diffusion system (5.13). This latter system have been intensively studied, in particular existence of traveling waves, describing propagation phenomena (see [84] or [230]). The main interest in this reduction process is that since the behavior of solutions to the scalar equation (5.13) is well-known. Therefore we can deduce, for small values of $\epsilon$, the local in time behavior of solutions to (5.4).

We observe that the limit reaction term $r(p) := pF_1(h(p),p)$ in (5.13) satisfies

$$r(0) = 0, \qquad r(1) = F_1(h(1),1) = 0,$$

because $H(h(p),p) = 0 = -pF_1(h(p),p) - (1-p)F_2(h(p),p)$. It means that the states $p^0 = 0$ (only population 2) and $p^0 = 1$ (only population 1) are equilibria for this system.

Moreover,

$$r'(0) = F_1(h(0),0)$$

and

$$r'(1) = h'(1)\partial_n F_1(h(1),1) + \partial_p F_1(h(1),1).$$

Hence under some direct sign assumptions on $F_1$ and $\partial_p F_1$, the equilibria 0 and 1 for $p$ can be made stable in the limit equation, if $r'(0)$ and $r'(1)$ are negative. In particular, in the example in Section 5.3, the function $r$ is bistable.

**Remark 5.2.** *The assumption $n^{init,\epsilon} - h(0)$ uniformly bounded with respect to $\epsilon$ in $L^2(\mathbb{R}^d)$ together with the uniform bound of $p^{init,\epsilon}$ in $L^2(\mathbb{R}^d)$ imply, thanks to Assumption 5.3, that $n^{init} - h(p^{init,\epsilon})$ is bounded in $L^2(\mathbb{R}^d)$, uniformly in $\epsilon > 0$. Indeed, $h$ is Lipschitz on $(0,1)$ and by the triangle inequality, we have $\|n^{init} - h(p^{init,\epsilon})\|_{L^2} \leq \|n^{init,\epsilon} - h(0)\|_{L^2} + \|h\|_{Lip}\|p^{init,\epsilon}\|_{L^2}$.*

**Remark 5.3.** *One might be interested by the effect of the local introduction of a variant into a population at equilibrium. In this situation, at the time of introduction, variant 2 is at equilibrium whereas the introduction of variant 1 is modeled by a compactly supported continuous nonnegative function $\phi$. Then we have $n_2^{init,\epsilon} = \frac{1}{\epsilon} - h(0)$ on $\mathbb{R}^d \setminus supp\ \phi$, and we set $n_1^{init,\epsilon} = \phi n_2^{init,\epsilon}$. Then, $p^{init} = \frac{\phi}{1+\phi}$ and Theorem's assumption (5.11) boils down to assume that*

$$\frac{1}{\epsilon} - (1+\phi)n_2^{init,\epsilon} - h(0) \text{ is uniformly bounded with respect to } \epsilon \text{ in } L^\infty(\mathbb{R}^d).$$

Finally, we mention that we can relax the assumption (5.11) by assuming that the sequence $(p^{\text{init},\epsilon})_\epsilon$ is uniformly bounded with respect to $\epsilon$ in $L^2(\mathbb{R}^d)$ instead of assuming its convergence. In fact, we can extract a subsequence of $(p^{\text{init},\epsilon})_\epsilon$ that converges weakly towards $p^{\text{init}}$ and the result applies. But the uniqueness of the weak limit $p^{\text{init}}$ is not guaranteed and therefore, the result in Theorem 5.1 is available only up to an extraction of a subsequence.

## 5.3 Application to a biological example

### 5.3.1 Presentation of the model

We consider the case of *Wolbachia* in arthropod species (for the biology of this bacterium, see [233] ; for mathematical modeling, see [29], [83], [121], [50]). It is an endo-symbiont that is maternally transmitted, causes cytoplasmic incompatibility (CI), and has several other effects on its host. Here, we understand CI as a mechanism through which one of the possible crossings is less viable. More precisely, if an uninfected female is fertilized by an infected male, a fraction only of its eggs will eventually hatch and give birth to viable larvae. For more details about CI, we refer to [233]. In the case of *Aedes* mosquitoes, *Wolbachia* reduces lifespan, changes fecundity and blocks the development of dengue virus (see [172], [232], [128]). It is then a potential biological tool to fight dengue epidemics. However, it does not change the way mosquitoes move. Therefore, in order to model a *Wolbachia* invasion (assessed in the field in [118]) we are precisely in our setting. Several (two) *variants of the same species* interact with each other in a complex way.

Specifically, we define the uninfected death rate $d_u$. This rate is multiplied by $\delta > 1$ for infected mosquitoes: $d_i = \delta d_u$. We also define an uninfected fecundity $F_u$ for uninfected mosquitoes, $F_i = (1-s_f)F_u$ for infected mosquitoes ; a resource parameter $\sigma$ ; and a CI parameter $0 < s_h \leq 1$, which means that a fraction $s_h$ of uninfected females' eggs fertilized by infected males won't hatch. Parameters $\delta$, $s_f$ and $s_h$ have been estimated in several cases and can be found in the literature (see [29] and references therein). We will always assume $s_h > s_f$. (In practice, we usually have $s_f$ close to 0 and $s_h$ close to 1). Let us denote $n_i(t,x)$, resp. $n_u(t,x)$, the density of the infected, resp. uninfected, mosquitoes at time $t \geq 0$, position $x \in \mathbb{R}^d$.

Several models have been written, using these parameters. In [50] (if we ignore the drift speed $v \in \mathbb{R}^d$ they used, which amounts at a change of coordinates) one find

$$\begin{cases} \partial_t n_i - \nabla \cdot (A(x)\nabla n_i) = n_i(1 - \sigma(n_u + n_i)) - d_u n_i, \\ \partial_t n_u - \nabla \cdot (A(x)\nabla n_u) = n_u F_u (1 - s_h \frac{n_i}{n_u + n_i})(1 - \sigma(n_u + n_i)) - d_u n_u. \end{cases} \tag{5.14}$$

In this model, $\delta = 1$ and variables are scaled so that $F_u(1-s_f) = F_i = 1$. Here the reduced population is defined by $n = \frac{1}{\sigma} - (n_i + n_u)$. The corresponding dynamics in $(n,p)$ for (5.14) is written

$$\begin{cases} \partial_t n - \nabla\cdot(A(x)\nabla n) = \big(\sigma n(p + F_u(1-p)(1-s_h p)) - d_u\big), \\ \partial_t p - \nabla\cdot(A(x)\nabla p) + 2\frac{\nabla n}{n}A(x)\nabla p = \sigma np(1-p)\big(d_u - F_u(1-s_h p)\big), \end{cases} \tag{5.15}$$

In (5.15), the reaction term for $p$ depends on $n$ merely for its intensity (it is a multiplicative factor). In particular, the unstable steady state (defining, in some sense, *a* possible "threshold for invasion") is equal to $\frac{1}{s_h}(1-\frac{d_u}{F_u})$ does not depend on $n$.

To further reduce this class of models and prove the convergence towards (5.3), we introduce the parameter $\epsilon$ to characterize the high fertility and competition that result in a carrying capacity of order $\frac{1}{\epsilon}$. Then we propose the following generalization of (5.14), which incorporates also the different death rate and the reduction of fecundity,

$$\begin{cases} \partial_t n_i - \nabla\cdot(A(x)\nabla n_i) &= (1-s_f)F_u n_i\big(\frac{1}{\epsilon} - \sigma(n_i+n_u)\big) - \delta d_u n_i, \\ \partial_t n_u - \nabla\cdot(A(x)\nabla n_u) &= F_u n_u(1 - s_h\frac{n_i}{n_i+n_u})\big(\frac{1}{\epsilon} - \sigma(n_i+n_u)\big) - d_u n_u, \end{cases} \tag{5.16}$$

Straightforwardly, we can compute the equilibria for the associated dynamical system.

**Lemma 5.2.** *As soon as $s_f + \delta - 1 < \delta s_h$, there are four distinct equilibria associated with (5.16) in the non-negative quadrant.*

- *Wolbachia invasion steady state $(n^*_{iW}, n^*_{uW}) := (\frac{1}{\sigma\epsilon} - \frac{d_u}{F_u}\frac{\delta}{1-s_f}, 0)$ is stable;*
- *Wolbachia extinction steady state $(n^*_{iE}, n^*_{uE}) := (0, \frac{1}{\sigma\epsilon} - \frac{d_u}{F_u})$ is stable;*
- *The co-existence steady state is unstable and reads*

$$(n^*_{iC}, n^*_{uC}) := \big(\big(\frac{1}{\sigma\epsilon} - \frac{d_u}{F_u}\frac{\delta}{1-s_f}\big)\frac{\delta-(1-s_f)}{\delta s_h}, \big(\frac{1}{\sigma\epsilon} - \frac{d_u}{F_u}\frac{\delta}{1-s_f}\big)\frac{\delta(s_h-1)+(1-s_f)}{\delta s_h}\big);$$

- *The steady state $(0,0)$ is unstable.*

### 5.3.2 Large population asymptotic

We perform the limit $\epsilon \to 0$ for system (5.16). To recover notations from Theorem 5.1, we identify $n_1 = n_i$, $n_2 = n_u$. As above we define the reduced quantity $n = \frac{1}{\sigma\epsilon} - (n_1+n_2)$ and $p = \frac{n_1}{n_1+n_2}$. Then with the notations in Section 5.2, one has

$$\begin{aligned} F_1(n,p) &= \sigma n(1-s_f)F_u - \delta d_u, \\ F_2(n,p) &= \sigma n F_u(1-s_h p) - d_u. \end{aligned}$$

Therefore, by definition (5.8), we compute

$$\begin{aligned} H(n,p) &= -p(\sigma n(1-s_f)F_u - \delta d_u) - (1-p)(\sigma n F_u(1-s_h p) - d_u) \\ &= -\sigma F_u n(s_h p^2 - (s_f+s_h)p + 1) + d_u((\delta-1)p+1). \end{aligned}$$

And Assumption 5.2 is satisfied. Then, Assumption 5.3 is easy to check since $H(0,p) = d_u((\delta-1)p+1) > 0$ and using the fact that the polynomial $x \mapsto s_h x^2 - (s_f+s_h)x + 1$ is minimal for $x = \frac{s_f+s_h}{2s_h}$, we have

$$\partial_n H(n,p) = -\sigma F_u(s_h p^2 - (s_f+s_h)p + 1) \le -\sigma F_u(1 - \frac{(s_f+s_h)^2}{4s_h}) < 0,$$

since we have $s_f < s_h$. We notice also that this computation implies that the above second order polynomial in $p$ is always away from 0 on $[0,1]$. Moreover, recalling the definition $H(n,p) = 0$ if and only if $n = h(p)$ from Assumption 5.3, we can compute $h(p) = \frac{d_u}{\sigma F_u}\frac{(\delta-1)p+1}{s_h p^2 - (s_f+s_h)p+1}$. Under Assumption 5.1 on $A$, Theorem 5.1 applies, and $p^\epsilon$ converges towards the solution of the following equation

$$\begin{cases} \partial_t p^0 - \nabla\cdot(A(x)\nabla p^0) &= r(p^0), \\ p^0(t=0) &= p^{\text{init}}, \end{cases} \tag{5.17}$$

and the reaction term writes

$$r(p) = \delta d_u s_h \frac{p(1-p)(p-\theta)}{s_h p^2 - (s_f + s_h)p + 1}, \qquad \theta = \frac{s_f + \delta - 1}{\delta s_h},$$

which is bistable provided $\delta$ satisfies the condition from Lemma 5.2:

$$s_f + \delta - 1 < \delta s_h. \tag{5.18}$$

If $\delta = 1$, we find the ubiquitous value $\theta(= p^*) = \frac{s_f}{s_h}$, which corresponds to the model of spacial spread of Wolbachia proposed in [29]. In addition, this expression is coherent with the one in [203] for general $\delta$. Even though the equation for $p$ has already been suggested for a while, as far as we know, no convergence result as ours had been proved before from a two-populations model to the bistable equation.

A direct application of Theorem 5.1 establishes that, in the limit $\epsilon \to 0$, the derivation in [29] holds true in a strong topology.

**Corollary 5.1.** *Assume that $A$ satisfies Assumption 5.1. Given $n_1^{init,\epsilon}$ and $n_2^{init,\epsilon}$ such that there exists $p^{init} \in L^2(\mathbb{R}^d)$ such that $p^{init,\epsilon} \rightharpoonup p^{init}$ as $\epsilon \to 0$ in $L^2(\mathbb{R}^d)$-weak and $\frac{1}{\sigma\epsilon} - n_1^{init,\epsilon} - n_2^{init,\epsilon} - \frac{d_u}{\sigma F_u} \in L^2 \cap L^\infty(\mathbb{R}^d)$ with uniform bounds in $\epsilon > 0$, then Theorem 5.1 applies and the solutions $(n_i^\epsilon, n_u^\epsilon)_{\epsilon>0}$ of (5.16) satisfy the convergence result in (5.12) where the limiting equation is given in (5.17).*

### 5.3.3 Numerical illustration

A numerical illustration of this convergence result is shown in Figure 5.1. Parameters are fixed according to biologically relevant data (freely adapted from [88]). Time unit is the day, and parameters per day are $F_i = F_u = 1.12$ (hence $s_f = 0$), $d_u = 0.27$ and $d_i = 0.3$, then $\delta = \frac{d_i}{d_u} = \frac{10}{9}$. We choose $s_f = 0.1$ and $s_h = 0.8$. We take $\sigma = 1$, and $A(x) \equiv 0.1$, which amounts at choosing a space scale.

We discretize the one-dimensional computational domain $[-15; 15]$ with space step $\Delta x = 0.05$ and take a time step $\Delta t = 0.005$. The reaction diffusion equations are discretized thanks to semi-implicit finite difference scheme, the diffusion operator being treated implicitly (to avoid too restrictive stability conditions), while the reaction term is treated explicitly. Curves are plotted every 5000 iterations, at times (in days) $T_1 = 25$, $T_2 = 50$, $T_3 = 75$, $T_4 = 100$ and $T_5 = 125$. We display 4 numerical tests with the same initial data $p^{\text{init}}$ compactly supported, plotted in pluses (+). The blue lines represent the solution of the limiting system (5.17). In dashed red lines are plotted the numerical results for the system of two populations (5.16). We observe that the solution of the limiting bistable system (5.17) exhibits a traveling front which propagates into the whole domain. Then the numerical results for 4 different values of the parameter $\epsilon$ are represented. For large populations, we observe that as $\epsilon$ goes to 0 (recall that the order of magnitude of the population size is $\frac{1}{\sigma\epsilon}$), the solution to the whole system (5.16) gets closer to the one of the limiting system. However, for $\epsilon = 0.6$, the introduced population goes extinct, and $p$ does not converge towards a traveling wave. This illustrates how the 2 by 2 system qualitatively differs from the limit reaction-diffusion equation.

An additional conclusion we can draw from Figure 5.1 is that our approximation result will always be *local in time*. Indeed, for small $\epsilon$ we see a traveling wave appear in dashed red, that has a *slower* speed than the blue one. Hence the norm of their difference will be constantly growing in time.

## 5.4 Proof of convergence

This Section is devoted to the proof of Theorem 5.1. We write the system of equations satisfied by $(n^\epsilon, p^\epsilon)$

$$\begin{cases} \partial_t n^\epsilon - \nabla \cdot (A(x)\nabla n^\epsilon) = \left(\frac{1}{\epsilon} - n^\epsilon\right) H(n^\epsilon, p^\epsilon), \\ \partial_t p^\epsilon - \nabla \cdot (A(x)\nabla p^\epsilon) + 2\epsilon \nabla p^\epsilon \cdot \frac{A(x)\nabla n^\epsilon}{1-\epsilon n^\epsilon} = p^\epsilon(1-p^\epsilon)\big(F_1 - F_2\big)(n^\epsilon, p^\epsilon), \\ n^\epsilon(t=0) = n^{\text{init},\epsilon}, \qquad p^\epsilon(t=0) = p^{\text{init},\epsilon}. \end{cases} \tag{5.19}$$

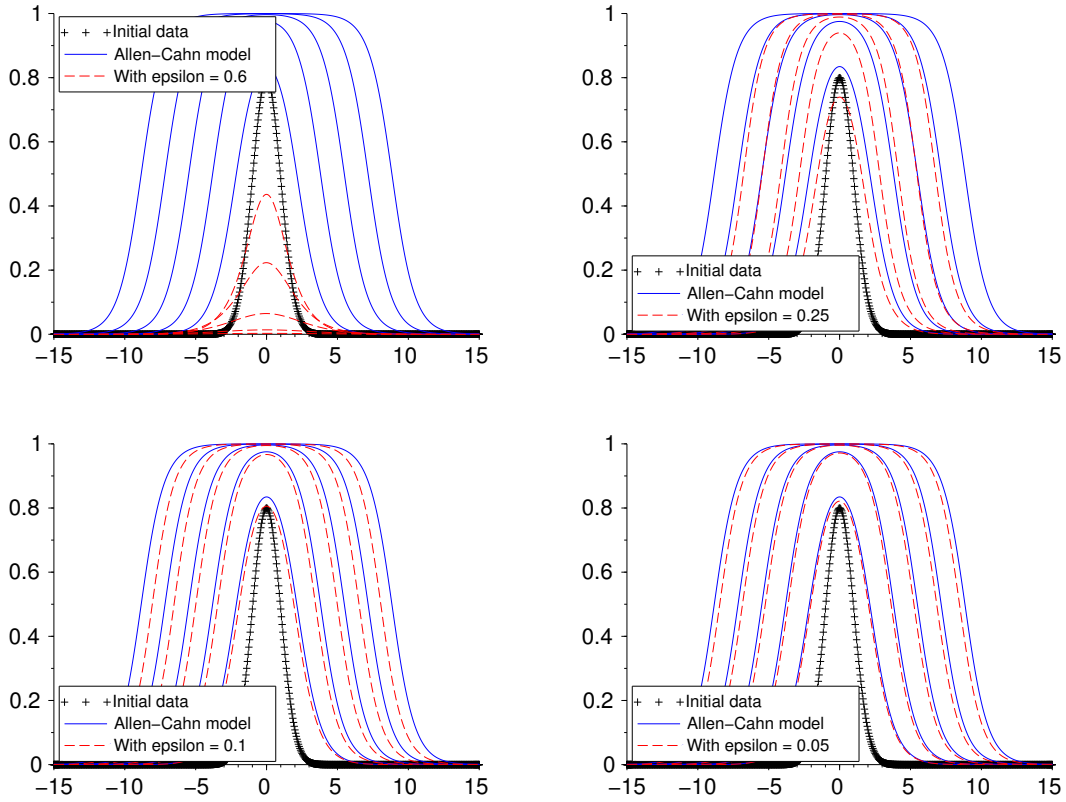



Figure 5.1: Initial data (+) creating a traveling wave in the limit system (blue) and convergence of the two-population solution (dashed red) as $\epsilon$ diminishes.

We recall that the initial data are assumed to satisfy (5.11). Then the sequence $(p^{\text{init},\epsilon})_\epsilon$ is bounded uniformly in $\epsilon$ in $L^2(\mathbb{R}^d)$ and $(n^{\text{init},\epsilon} - h(0))_\epsilon$ is bounded uniformly in $\epsilon$ in $L^2 \cap L^\infty(\mathbb{R}^d)$. The proof of Theorem 5.1 relies strongly on a sequence of a priori estimates uniform in $\epsilon$, which give compactness and allow to pass to the limit in the equation for $p^\epsilon$.

From now on, we will drop the superscript $\epsilon$ in the notations.

### 5.4.1 Estimates

For $\epsilon > 0$ fixed, existence of solutions to (5.19) is classical (see e.g. [187]). Now we establish some a priori estimates uniform in $\epsilon > 0$. First, we have the following $L^\infty$ bounds.

**Lemma 5.3.** *Under the assumptions of Theorem 5.1, for any positive initial data, the unique solution $(p, n)$ to (5.19) satisfies*

$$\forall t > 0, x \in \mathbb{R}^d,\ 0 \le p(t,x) \le 1$$

*and $n \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$. Moreover, there exists $\epsilon_0 > 0$ such that the $L^\infty$ bound on $n$ is uniform in $\epsilon_0 > \epsilon > 0$.*

*Proof.* As stated before, positivity of $n_1, n_2$ is straightforward and implies the uniform bounds on $p$ in $L^\infty$.

Using Stampacchia's method for the bound on $n$, we notice that, from Assumption 5.3, for all $p \in [0,1]$, $\left(\frac{1}{\epsilon} - n\right) H(n,p)$ is positive for $n$ between 0 and $h(p)$ and negative afterwards until $\frac{1}{\epsilon}$. Then, for $\tilde{K} = \max_{p\in[0,1]} h(p)$, we define $y(t) = \int_{\mathbb{R}^d} \left(n(t,x) - \tilde{K}\right)_+ dx$. Multiplying the equation on $n - \tilde{K}$ by $1_{n>\tilde{K}}$ and integrating over $\mathbb{R}^d$ gives, for $\epsilon < \frac{1}{\tilde{K}}$

$$\frac{d}{dt} y(t) + \int_{\mathbb{R}^d} \nabla (n - \tilde{K})_+ \cdot A(x) \nabla (n - \tilde{K})_+ dx \le 0.$$

And in particular, $\frac{d}{dt}y < 0$. Since, from Assumption (5.11), $n^{\text{init}}$ is bounded in $L^\infty$ uniformly with respect to $\epsilon$, we can pick $\tilde{K}$ such that $\tilde{K} > \|n^{\text{init}}\|_\infty$. Then $y(0) = 0$. We deduce that $y \equiv 0$.

To conclude, the result is proved with

$$\epsilon_0 = \big(\max\big(\max_{p\in[0,1]} h(p), \|n^{\text{init}}\|_\infty\big)\big)^{-1}.$$

□

Now, we aim at getting the following boundedness result.

**Proposition 5.1.** *Let $T > 0$. Under the assumptions in Theorem 5.1, we define $M := n - h(p)$. Then, there exists $\epsilon_0 > 0$ such that $M$ and $p$ are uniformly bounded in $\mathcal{H}^1_T \cap \mathcal{H}^2_T$, for all $\epsilon \le \epsilon_0$.*

We recall that the function $h$ is defined in Assumption 5.3 and belongs to $\mathcal{C}^2([0,1])$. Then we may define

$$h_0 = \|h\|_{L^\infty([0,1])}, \quad h_0' = \|h'\|_{L^\infty([0,1])}, \quad h_0'' = \|h''\|_{L^\infty([0,1])}. \tag{5.20}$$

We notice that, by definition and from Lemma 5.3, we have that $M$ is uniformly bounded in $L^\infty$ for $\epsilon \le \epsilon_0$. The proof of this result relies on estimates on $p$ and $M$, and we postpone the proof of Proposition 5.1 after proving them in the two following technical Lemma. The first one is for $p$.

**Lemma 5.4.** *There is a positive constant $K$ independent of $\epsilon$ such that $\forall \epsilon > 0$,*

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d} p^2 dx + (1 - \epsilon C_1)\int_{\mathbb{R}^d} \nabla p A(x)\nabla p dx \le \epsilon C_2 \int_{\mathbb{R}^d} \nabla M A(x)\nabla M dx + K\int_{\mathbb{R}^d} p^2 dx,$$

*where $C_1 = 2\big(1 + \frac{h_0'}{2} + (h_0')^2\big)$ and $C_2 = 2(1 + \frac{h_0'}{2})$.*

*Proof.* We multiply by $p$ the equation satisfied by $p$ in (5.19), and integrate over $\mathbb{R}^d$

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d} p^2 dx + \int_{\mathbb{R}^d} \nabla p A(x)\nabla p dx + 2\epsilon\int_{\mathbb{R}^d} \frac{p}{1-\epsilon n}\nabla p \cdot A(x)\nabla n dx \\ \le \int_{\mathbb{R}^d} p^2(1-p)\big(F_1 - F_2\big)(n,p)dx. \tag{5.21}$$

Thanks to Lemma 5.3, we know that $\frac{p}{1-\epsilon n}$ is well-defined for $\epsilon$ small enough, and the denominator is uniformly positive. Hence we may use a Cauchy-Schwarz inequality,

$$\int_{\mathbb{R}^d} \frac{p}{1-\epsilon n}\nabla n A(x)\nabla p dx \le \frac{1}{2}\int_{\mathbb{R}^d} \frac{p}{1-\epsilon n}\nabla p A(x)\nabla p dx + \frac{1}{2}\int_{\mathbb{R}^d} \frac{p}{1-\epsilon n}\nabla n A(x)\nabla n dx.$$

Since $n = M + h(p)$, we have $\nabla n = \nabla M + h'(p)\nabla p$. We may also write,

$$\int_{\mathbb{R}^d} \nabla n A(x)\nabla n dx \le ((h_0')^2 + \frac{h_0'}{2})\int_{\mathbb{R}^d} \nabla p A(x)\nabla p dx + (1 + \frac{h_0'}{2})\int_{\mathbb{R}^d} \nabla M A(x)\nabla M dx.$$

Now, collecting these inequalities for $\epsilon$ small enough, such that $\frac{p}{1-\epsilon n} \le 2$ yields

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d} p^2 dx + \Big(1 - 2\epsilon\big(1 + h_0'(h_0' + \frac{1}{2})\big)\Big)\int_{\mathbb{R}^d} \nabla p A(x)\nabla p \\ \le 2\epsilon\big(1 + \frac{h_0'}{2}\big)\int_{\mathbb{R}^d} \nabla M A(x)\nabla M + K_F\int_{\mathbb{R}^d} p^2 dx,$$

where

$$K_F := \sup\{|(1-p)(F_1 - F_2)(n,p)|, \text{ as } |n| \leqslant \sup_{0<\epsilon\leqslant\epsilon_0} \|n\|_{L^\infty} \text{ and } 0 \leqslant p \leqslant 1\}. \tag{5.22}$$

Thanks to Lemma 5.3 and the continuity of the functions $F_1$ and $F_2$, the constant $K_F$ is finite. This is the expected estimate. □

Similarly, on $M := n - h(p)$,

**Lemma 5.5.** *There are positive constants $C_3, C_4, C_5, C_6$ independent of $\epsilon$ such that, for all $\epsilon > 0$,*

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d} M^2 dx + (1-\epsilon C_3)\int_{\mathbb{R}^d} \nabla M A(x)\nabla M dx \\ \leq \big(C_4 - \frac{C_5}{\epsilon}\big)\int_{\mathbb{R}^d} M^2 dx + C_6\big(\int_{\mathbb{R}^d} p^2 dx + \int_{\mathbb{R}^d} \nabla p A(x)\nabla p dx\big),$$

*where $C_3 = \frac{h_0'}{2}$, $C_4 = B(h_0 + \tilde{K}') + \frac{h_0' K_F}{2}$, $C_5 = B$ and $C_6 = \max\big(\frac{h_0' K_F}{2}, \tilde{K}' h_0'' + \epsilon h_0'(h_0' + \frac{1}{2})\big)$, and $\tilde{K}', K_F$ are positive constants defined in (5.22) and (5.26).*

*Proof.* The quantity $M$ satisfies the following equation (obtained from (5.19))

$$\begin{aligned}\partial_t M - \nabla\cdot(A(x)\nabla M) &= \partial_t n - \nabla\cdot(A(x)\nabla n) - h'(p)\partial_t p + \nabla\cdot(h'(p)A(x)\nabla p)\\ &= \big(\frac{1}{\epsilon} - M - h(p)\big)H(M+h(p),p) + h''(p)\nabla p\cdot(A(x)\nabla p)\\ &\quad - h'(p)\Big((1-p)(-F_2 - H)(M+h(p),p)\\ &\quad - 2\epsilon\frac{1}{1-\epsilon n}(A(x)\nabla M + h'(p)A(x)\nabla p)\cdot\nabla p\Big), \end{aligned} \tag{5.23}$$

it is associated with an initial data $M^{\text{init}} = n^{\text{init}} - h(p^{\text{init}})$ bounded in $L^2(\mathbb{R}^d)$. Indeed, as noted in Remark 5.3, we have,

$$|n^{\text{init}} - h(p^{\text{init}})| \leq |n^{\text{init}} - h(0)| + |h(0) - h(p^{\text{init}})| \leq |n^{\text{init}} - h(0)| + h_0'|p^{\text{init}}|.$$

Moreover, from (5.11), $p^{\text{init}}$ is bounded in $L^2(\mathbb{R}^d)$ and $n^{\text{init}} - h(0)$ is bounded in $L^2(\mathbb{R}^d)$, with uniform bounds in $\epsilon$. It implies the uniform bound of $M^{\text{init}}$ in $L^2(\mathbb{R}^d)$.

Now, we assume that $\epsilon$ is small enough, so that the term $\frac{1}{\epsilon} - M - h(p)$ remains positive (this is possible thanks to Lemma 5.3). We multiply by $M$ equation (5.23), and integrate over $\mathbb{R}^d$

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d} M^2 dx + \int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx &\\ \leq -B\int_{\mathbb{R}^d} M^2\big(\frac{1}{\epsilon} - M - h(p)\big)dx &\\ + \int_{\mathbb{R}^d} 2Mh'(p)\epsilon\frac{1}{1-\epsilon n}(A(x)\nabla M + h'(p)A(x)\nabla p)\cdot\nabla p dx &\\ + \int_{\mathbb{R}^d} Mh''(p)\nabla p\cdot(A(x)\nabla p)dx &\\ - \int_{\mathbb{R}^d} Mh'(p)(1-p)(-F_2 - H)(M+h(p),p)dx &\end{aligned} \tag{5.24}$$

Since $\partial_n H \leq -B$ (Assumption 5.3), $\dfrac{H(M+h(p),p) - H(h(p),p)}{M} \leq -B$. Multiplying this inequality by $M^2 \geqslant 0$ we get $MH(M+h(p)) \leq -BM^2$ because $H(h(p),p) = 0$ for all $p \in [0,1]$.

Now, we bound each one of these terms of the right hand side of (5.24) separately (keeping in mind the fact that $\epsilon$ will be chosen small enough)

$$-B\int_{\mathbb{R}^d} M^2\big(\frac{1}{\epsilon} - M - h(p)\big)dx \leq -B\big(\frac{1}{\epsilon} - h_0 - \tilde{K}'\big)\int_{\mathbb{R}^d} M^2 dx, \tag{5.25}$$

where

$$\tilde{K}' = \sup\{|n - h(p)|,\ \text{as } |n| \leq \sup_{0<\epsilon\leqslant\epsilon_0}\|n\|_{L^\infty} \text{ and } 0 \leqslant p \leqslant 1\}. \tag{5.26}$$

From Lemma 5.3, $\tilde{K}'$ is finite and by definition $M = n - h(p)$, $\tilde{K}'$ bounds $|M|$. We pick $\epsilon < \epsilon_0$ such that $1 - \epsilon n > \frac{1}{2}$ (again, using Lemma 5.3). After using a Cauchy-Schwarz inequality, we get

$$\Big|\int_{\mathbb{R}^d} 2Mh'(p)\epsilon\frac{1}{1-\epsilon n}(A(x)\nabla M + h'(p)A(x)\nabla p)\cdot\nabla p dx\Big| \\ \leqslant h_0'\tilde{K}'\epsilon\Big((h_0' + \frac{1}{2})\int_{\mathbb{R}^d}\nabla p\cdot(A(x)\nabla p)dx + \frac{1}{2}\int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx\Big). \tag{5.27}$$

Finally, by definition of $H$ in (5.8), we have

$$(-F_2 - H)(n,p) = p(F_1 - F_2)(n,p).$$

Then using the constant $K_F$ defined in (5.22), we deduce, applying a Cauchy-Schwarz inequality,

$$\left|\int_{\mathbb{R}^d} \big(Mh''(p)\nabla p\cdot(A(x)\nabla p) - Mh'(p)(1-p)(-F_2-H)(M+h(p),p)\big)dx\right| \leqslant h_0''\tilde{K}'\int_{\mathbb{R}^d}\nabla p\cdot(A(x)\nabla p)dx + \frac{h_0'K_F}{2}\int_{\mathbb{R}^d}M^2dx + \frac{h_0'K_F}{2}\int_{\mathbb{R}^d}p^2dx. \quad (5.28)$$

Combining (5.25), (5.27) and (5.28) we get

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}M^2dx + \int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx &\leqslant -B(\frac{1}{\epsilon} - h_0 - \tilde{K}' - \frac{h_0'K_F}{2B})\int_{\mathbb{R}^d}M^2dx + \frac{h_0'K_F}{2}\int_{\mathbb{R}^d}p^2dx\\ &\quad + \big(\tilde{K}'h_0'' + h_0'\epsilon(h_0' + \frac{1}{2})\big)\int_{\mathbb{R}^d}\nabla p\cdot(A(x)\nabla p)dx\\ &\quad + h_0'\frac{\epsilon}{2}\int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx.\end{aligned}$$

This is the expected estimate. □

With Lemmas 5.4 and 5.5 we can proceed to prove Proposition 5.1.

*Proof of Proposition 5.1.* Let $\alpha > 0$, summing the inequality in Lemmas 5.4 and 5.5, we obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}(M^2+\alpha p^2)dx + (1-\epsilon C_3 - \alpha\epsilon C_2)\int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx\\ &\quad + (\alpha(1-\epsilon C_1) - C_6)\int_{\mathbb{R}^d}\nabla p\cdot(A(x)\nabla p)dx \leq \big(C_4 - \frac{C_5}{\epsilon}\big)\int_{\mathbb{R}^d}M^2dx + \big(C_6+\alpha K\big)\int_{\mathbb{R}^d}p^2dx.\end{aligned}$$

Now, we can pick $\alpha > 0$, $\epsilon_0' \in (0,\epsilon_0)$ such that for all $0 < \epsilon < \epsilon_0'$,

$$1 - \epsilon(C_3 + \alpha C_2) \geq \frac{1}{2}, \quad \text{and} \quad \alpha(1-\epsilon C_1) - C_6 \geq \frac{1}{2}.$$

The choice $\alpha = C_6 + 1$, then $\epsilon_0' = \min(\frac{1}{2(C_3+C_2(C_6+1))}, \frac{1}{2C_1(C_6+1)})$ suffices. Hence we arrive at

$$\begin{aligned}&\frac{d}{dt}\int_{\mathbb{R}^d}(M^2+\alpha p^2)dx + \int_{\mathbb{R}^d}\nabla M\cdot(A(x)\nabla M)dx + \int_{\mathbb{R}^d}\nabla p\cdot(A(x)\nabla p)dx\\ &\qquad \leq 2\big(C_4 - \frac{C_5}{\epsilon}\big)\int_{\mathbb{R}^d}M^2dx + 2\big(C_6+\alpha K\big)\int_{\mathbb{R}^d}p^2dx. \quad (5.29)\end{aligned}$$

Next, using the positivity of $A$, we may write for all $\epsilon > 0$ smaller than $\epsilon_0$ and $C_4/C_5$

$$\frac{d}{dt}\int_{\mathbb{R}^d}(M^2+\alpha p^2)dx \leq 2\frac{C_6+\alpha K}{\alpha}\int_{\mathbb{R}^d}(\alpha p^2 + M^2)dx,$$

and thus by Gronwall's lemma, for all $\epsilon > 0$ small enough, with $C_0 := 2\frac{C_6+\alpha K}{\alpha}$

$$\int_{\mathbb{R}^d}\big(M^2(t,x) + \alpha p^2(t,x)\big)dx \leq e^{C_0 t}\big(\|M^{\text{init}}\|^2_{L^2(\mathbb{R}^d)} + \alpha\|p^{\text{init}}\|^2_{L^2(\mathbb{R}^d)}\big).$$

Since initial data are uniformly bounded in $L^2(\mathbb{R}^d)$ thanks to (5.11) and Remark 5.2, the first part of Proposition 5.1 is proved. For all $T > 0$, $M$ and $p$ are uniformly bounded in $\mathcal{H}^1_T$ for $\epsilon$ small enough.

The second part follows easily from a time integration of (5.29). If $\epsilon$ is small enough, we get

$$\begin{aligned}\int_0^t\int_{\mathbb{R}^d}\big(\nabla M\cdot(A(x)\nabla M) + \nabla p\cdot(A(x)\nabla p)\big)dxds &\leq 2\big(C_6+\alpha K\big)\int_0^t\int_{\mathbb{R}^d}p^2dxds\\ &\quad + \int_{\mathbb{R}^d}\big((M^{\text{init}})^2 + \alpha(p^{\text{init}})^2\big)dx.\end{aligned}$$

Since we have proved the uniform $L^2$-bound of $p$, we conclude from the positivity of $A$ (Assumption 5.1), and the uniform bounds on the initial data. □

### 5.4.2 Convergence of $M$

Until now we have not used the strength of the negative term in $\frac{1}{\epsilon}$ in the right hand side of (5.29). Thanks to it, we can even get convergence of $M$.

**Lemma 5.6.** *Under the assumptions of Theorem 5.1, for all $T > 0$, $M \xrightarrow[\epsilon\to 0]{} 0$ strongly in $\mathcal{H}^1_T$.*

*Proof.* Back to the estimate in Lemma 5.5, and thanks to Proposition 5.1, we may write

$$\frac{d}{dt}\int_{\mathbb{R}^d} M^2 dx \le \big(2C_4 - \frac{2C_5}{\epsilon}\big)\int_{\mathbb{R}^d} M^2 dx + C(t),$$

where $C(t) := 2C_6\big(\int_{\mathbb{R}^d} p^2 dx + \int_{\mathbb{R}^d} \nabla p\cdot(A(x)\nabla p)dx\big)$. From Proposition 5.1, we deduce that $C$ is bounded in $L^1(0,T)$. Applying a Gronwall's lemma, we may write

$$\int_{\mathbb{R}^d} M^2 dx \le e^{-2(\frac{C_5}{\epsilon}-C_4)t}\big(\|M^{\text{init}}\|_{L^2(\mathbb{R}^d)} + \int_0^t e^{2(\frac{C_5}{\epsilon}-C_4)t'}C(t')dt'\big).$$

Let $\epsilon$ be small enough such that $\frac{C_5}{\epsilon} > C_4$. Then integrating the latter inequality for $t\in[0,T]$, we deduce

$$\int_0^T\int_{\mathbb{R}^d} M^2 dxdt \le \frac{\epsilon}{2(C_5-\epsilon C_4)}\|M^{\text{init}}\|_{L^2(\mathbb{R}^d)} + \int_0^T\int_0^t e^{2(\frac{C_5}{\epsilon}-C_4)(t'-t)}C(t')dt'.$$

We make a change of variable to estimate the last term in the right hand side:

$$\begin{aligned}\int_0^T\int_0^t e^{2(\frac{C_5}{\epsilon}-C_4)(t'-t)}C(t')dt'dt &= \int_0^T\int_{t'}^T e^{2(\frac{C_5}{\epsilon}-C_4)(t'-t)}dt\,C(t')dt'\\ &= \int_0^T\int_{t'-T}^0 e^{2(\frac{C_5}{\epsilon}-C_4)\tau}d\tau\,C(t')dt'\\ &\le \frac{\epsilon}{2(C_5-\epsilon C_4)}\int_0^T C(t')dt'.\end{aligned}$$

We conclude that

$$\int_0^T\int_{\mathbb{R}^d} M^2 dxdt \le \frac{\epsilon}{2(C_5-\epsilon C_4)}\left(\|M^{\text{init}}\|_{L^2(\mathbb{R}^d)} + \int_0^T C(t')dt'\right).$$

It implies the expected convergence as $\epsilon\to 0$. □

### 5.4.3 Compactness result and proof of Theorem 5.1

Before proving our main result, we recall the following compactness result (see [205]).

**Lemma** (Lions-Aubin)**.** *Let $T>0, q\in(1,\infty)$, $(\psi_n)_n$ a bounded sequence in $L^q(0,T;H)$, where $H$ is a Banach space. If $\psi_n$ is bounded in $L^q(0,T;V)$ and $V$ compactly embeds in $H$, and if $(\partial_t\psi_n)_n$ is bounded in $L^q(0,T;V')$ uniformly with respect to $n$, then $(\psi_n)_n$ is relatively compact in $L^q(0,T;H)$.*

*Proof of Theorem 5.1.* We split the proof into three steps. First, our previous estimates together with Lions-Aubin lemma enable us to prove relative compactness on bounded domains. Then, through a diagonal extraction process, we prove that there exists (up to extracting a subsequence) a global limit. Finally, thanks to our uniform estimates, we prove that this limit satisfies a universal equation whose solution is unique, which in turn implies convergence of the whole sequence.

**Step 1: Local relative compactness.** For $R>0$ we define the increasing sequence $(B_R)_R$ of balls of radius $R$ with center 0 in $\mathbb{R}^d$, and $H_R = L^2(B_R)$, $V_R = H^1(B_R)\cap L^\infty(B_R)$, and pick $T>0$. Then, we check that Lions-Aubin Lemma with $q=2$ can be applied to

$$\psi_\epsilon^{(R)} = p^\epsilon|_{B_R}.$$

Lemma 5.4 gives boundedness in $L^q(0,T;V_R)$. The compact embedding is classical (Rellich-Kondrachov). We check that the time derivative is bounded. Let $\chi \in V_R$, $\langle \cdot\,,\cdot \rangle = \langle \cdot\,,\cdot \rangle_{V_R',V_R}$ and $t \in (0,T)$.

$$\int_0^t |\langle \partial_t p^\epsilon(\tau),\chi\rangle|^2 d\tau = \int_0^t \Big|\langle \nabla\cdot(A(x)\nabla p^\epsilon) - 2\epsilon \nabla p^\epsilon \cdot \frac{A(x)\nabla n^\epsilon}{1-\epsilon n^\epsilon} + p^\epsilon(1-p^\epsilon)(F_1-F_2)(n^\epsilon,p^\epsilon),\chi\rangle\Big|^2 d\tau.$$

This can be bounded

$$\begin{aligned}\int_0^t |\langle \partial_t p^\epsilon(\tau),\chi\rangle|^2 d\tau &\le \Big(\int_0^t\int_{B_R} |A(x)\nabla p^\epsilon\cdot\nabla\chi|\Big)^2 \\ &\quad + \epsilon\Big(\int_0^t\int_{B_R}|A(x)\nabla n^\epsilon\cdot\nabla p^\epsilon|\Big)^2 + Ct\int_{B_R}\chi^2 \\ &\le \|\nabla\chi\|_{H_R}^2 \|\nabla p^\epsilon\|_{L^2(0,T;H_R)}^2 \\ &\quad + 2\epsilon \|\nabla n^\epsilon\|_{L^2(0,T;H_R)}^2 \|\nabla p^\epsilon\|_{L^2(0,T;H_R)}^2 \|\chi\|_\infty^2 + CT\|\chi\|_{H_R}^2,\end{aligned}$$

which gives the required bound, uniform in $0<\epsilon<\epsilon_0$, for $\epsilon_0$ small enough. This holds thanks to Lemmas 5.4 and 5.5.

**Step 2: Global convergence.** Now, for all $R \in \mathbb{Z}_{>0}$, one can extract converging (in $L^2(0,T;H_R)$) subsequence from $(p^\epsilon)_\epsilon$ by Lions-Aubin Lemma. We perform a diagonal extraction process successively in $R$, so that

$$p^{\epsilon_m^{(R)}} \xrightarrow[m\to\infty]{} p^{(R)} \text{ in } L^2(0,T;H_R),$$

and by construction $(\epsilon_m^{(R_1)})_m$ is a subsequence of $(\epsilon_m^{(R_2)})_m$ if $R_2 > R_1$. Because the whole family $(p^\epsilon)_\epsilon$ is in $L^2(0,T;H^1(\mathbb{R}^d))$ uniformly in $\epsilon$ (by Lemma 5.4), one gets weak convergence of gradient

$$\nabla p^{\epsilon_m^{(R)}} \xrightarrow[m\to\infty]{} \nabla p^{(R)} \text{ in } L^2(0,T;H_R).$$

Thanks to Lemma 5.4, we know that the limits $p^{(R)}$ are well-defined, do not depend on the extracted subsequences, satisfy the same bounds as $(p^\epsilon)_\epsilon$ and

$$R_2 > R_1 \implies p^{(R_2)}\big|_{B_{R_1}} = p^{(R_1)}.$$

Therefore we can define $p^0 \in L^2(0,T;L^2(\mathbb{R}^d))$ and we have constructed a subsequence, still denoted $(p^\epsilon)_\epsilon$, such that $p^\epsilon \xrightarrow[\epsilon\to0]{} p^0$ strongly in $L^2(0,T;L^2(B_R))$ for all $R>0$.

To pass from local to global convergence, we need to have uniform in $\epsilon$ estimate in the tails $|x|>R$. To do so, let us introduce $\phi \in \mathcal{C}^\infty(\mathbb{R}^d)$ such that $0\le\phi\le1$, $\phi(x)=0$ if $|x|<1/2$ and $\phi(x)=1$ if $|x|>1$. Then we denote $\phi_R(x)=\phi(x/R)$. Multiplying the equation satisfied by $p^\epsilon$ in (5.19) by $p^\epsilon\phi_R$ and integrating over $\mathbb{R}^d$, we deduce

$$\begin{aligned}\frac12\frac{d}{dt}\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx + \int_{\mathbb{R}^d}\nabla(p^\epsilon\phi_R)\cdot A(x)\nabla p^\epsilon\,dx + \int_{\mathbb{R}^d}\frac{2\epsilon}{1-\epsilon n^\epsilon}\phi_R p^\epsilon\nabla p^\epsilon\cdot A(x)\nabla n^\epsilon\,dx \\ \le K_F\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx,\end{aligned}$$

where $K_F$ has been defined in (5.22). Using a Cauchy-Schwarz inequality, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\nabla(\phi_R p^\epsilon)\cdot A(x)\nabla p^\epsilon\,dx &= \int_{\mathbb{R}^d} p^\epsilon\nabla\phi_R\cdot A(x)\nabla p^\epsilon\,dx + \int_{\mathbb{R}^d}\phi_R\nabla p^\epsilon\cdot A(x)\nabla p^\epsilon\,dx \\ &\ge -\Big(\int_{\mathbb{R}^d}\nabla\phi_R\cdot A(x)\nabla\phi_R\,dx\Big)^{1/2}\Big(\int_{\mathbb{R}^d}\nabla p^\epsilon\cdot A(x)\nabla p^\epsilon\,dx\Big)^{1/2}.\end{aligned}$$

By definition of $\phi_R$ we have that $\nabla\phi_R(x)=\frac1R\nabla\phi(x/R)$ and $\nabla\phi_R(x)=0$ on $B_{R/2}\cup\mathbb{R}^d\setminus\overline{B_R}$. As above, we take $\epsilon$ small enough such that $1-\epsilon n^\epsilon \ge \frac12$, which can be done thanks to Lemma 5.3. Then, as in the proof of Proposition 5.1, there exists a nonnegative function $C(t)\in L^1(0,T)$ such that, thanks to a Cauchy-Schwarz inequality

$$\Big|\int_{\mathbb{R}^d}\frac{2\epsilon}{1-\epsilon n^\epsilon}\phi_R p^\epsilon\nabla p^\epsilon\cdot A(x)\nabla n^\epsilon\,dx\Big| \le C(t)\epsilon, \qquad \text{and} \qquad -\int_{\mathbb{R}^d}\nabla\phi_R\cdot A(x)\nabla p^\epsilon\,dx \le \frac{C(t)}{R}.$$

Then, we have obtained

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx \le C(t)\big(\frac{1}{R}+\epsilon\big)+K_F\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx.$$

Using a Gronwall Lemma, it implies

$$\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx \le e^{2K_F t}\int_{\mathbb{R}^d}(p^{\rm init})^2\phi_R\,dx+\big(\frac{1}{R}+\epsilon\big)\int_0^t 2C(\tau)e^{2K_F(t-\tau)}\,d\tau.$$

By definition of $\phi_R$ ($\phi_R(x)=1$ on $\mathbb{R}^d\setminus B_R$), we deduce that for all $\epsilon>0$ small enough and all $R>0$,

$$\begin{aligned}\int_0^T\int_{\mathbb{R}^d\setminus B_R}|p^\epsilon|^2\,dx &\le \int_0^T\int_{\mathbb{R}^d}(p^\epsilon)^2\phi_R\,dx\\ &\le \frac{e^{2K_F T}-1}{2K_F}\left(\int_{\mathbb{R}^d\setminus B_{R/2}}(p^{\rm init})^2\,dx+(\frac{1}{R}+\epsilon)\int_0^T 2C(t)\,dt\right).\end{aligned}\tag{5.30}$$

It implies a uniform bound, since $p^{\rm init}\in L^2(\mathbb{R}^d)$.

Finally, we conclude that the subsequence $(p^\epsilon)_\epsilon$ converges strongly towards $p^0$ in $L^2(0,T;L^2(\mathbb{R}^d))$ as $\epsilon\to 0$. Indeed, we have

$$\int_0^T\int_{\mathbb{R}^d}|p^\epsilon-p^0|^2\,dxdt=\int_0^T\int_{B_R}|p^\epsilon-p^0|^2\,dxdt+\int_0^T\int_{\mathbb{R}^d\setminus B_R}|p^\epsilon-p^0|^2\,dxdt.$$

The second term of the right hand side is uniformly bounded for $R$ large enough thanks to (5.30) and the fact that $p^0\in L^2(0,T;L^2(\mathbb{R}^d))$. For the first term we use the local convergence.

**Step 3: Limit equation.** From the strong convergence of the sequence $(p^\epsilon)\epsilon$ in $L^2(0,T;L^2(\mathbb{R}^d))$ and the Lipschitz continuity of the function $h$, we deduce that $(h(p^\epsilon))_\epsilon$ converges strongly in $L^2(0,T;L^2(\mathbb{R}^d))$ towards $h(p^0)$. Moreover, using the triangle inequality, we have

$$|n^\epsilon-h(p^0)|\le|n^\epsilon-h(p^\epsilon)|+h_0'|p^\epsilon-p^0|.$$

Applying Lemma 5.6, we deduce that

$$n^\epsilon\xrightarrow[\epsilon\to0]{}n^0:=h(p^0)\text{ strongly in }L^2(0,T;L^2(\mathbb{R}^d)).\tag{5.31}$$

Then, we obtain the equation satisfied by $p^0$ using the weak forms of the equations on $p^\epsilon$ in (5.19): for all $\chi\in\mathcal{C}_c^\infty(\mathbb{R}^d)$,

$$\begin{aligned}\int_{\mathbb{R}^d}p^\epsilon(T,x)\chi(x)dx-\underbrace{\int_{\mathbb{R}^d}p^{{\rm init},\epsilon}(x)\chi(x)dx}_{\text{weak convergence}}+\underbrace{\int_0^T\int_{\mathbb{R}^d}\nabla p^\epsilon(t,x)\cdot A(x)\nabla\chi(x)dxdt}_{\text{weak convergence}}\\ +2\epsilon\underbrace{\int_0^T\int_{\mathbb{R}^d}\chi(x)\nabla p^\epsilon(t,x)\cdot\frac{A(x)\nabla n^\epsilon(t,x)}{1-\epsilon n^\epsilon(t,x)}dxdt}_{\text{bounded as }\epsilon\to0}\\ =\int_0^T\int_{\mathbb{R}^d}\chi(x)\underbrace{p^\epsilon(1-p^\epsilon)(F_1-F_2)(n^\epsilon,p^\epsilon)}_{\text{strong convergence}}dxdt\end{aligned}$$

We can pass to the limit in each term, using also (5.11) for the second term.

Hence $p^0$ is in $L^2(0,T;H^1(\mathbb{R}^d))$ and is a weak solution of the initial value problem

$$\begin{cases}\partial_t p^0-\nabla\cdot(A(x)\nabla p^0=p^0(1-p^0)(F_1-F_2)(n^0,p^0),\\ p^0(t=0,\cdot)=p^{\rm init}.\end{cases}\tag{5.32}$$

Using (5.31) in (5.32) yields a self-contained initial valued reaction-diffusion system on $p^0$ that has a unique solution. It defines in turn uniquely $n^0$ through (5.31). Since solutions to the initial value system (5.32) are unique, all extracted subsequences converge to the same limit. Therefore, the whole sequences converge, strongly in $L^2$ with weak convergence of gradients.

This concludes the proof of Theorem 5.1.

□

## 5.5 Generalization of the result

We have stated Theorem 5.1 so as to keep simplicity and stick to the biological application in Section 5.3. It can be slightly generalized in order to encompass spontaneous transition between variants.

Individuals in state 1 may give birth to individuals in state 2, and *vice versa*. To do so, we consider more general reaction term and replace system (5.1) by

$$\begin{cases} \partial_t n_1 - \nabla\cdot(A(x)\nabla n_1) &= \widetilde{f}_1(n_1,n_2), \\ \partial_t n_2 - \nabla\cdot(A(x)\nabla n_2) &= \widetilde{f}_2(n_1,n_2), \end{cases}$$

In fact, the basic property we require in our proof is that $p$ stays between 0 and 1, that is, $n_i$ remain non-negative. Here is the minimal hypothesis ensuring positivity (in the spirit of [187]).

**Assumption 5.4** (Positivity)**.** *We assume*

$$\forall n_1,n_2\in\mathbb{R}_+,\quad \widetilde{f}_1(0,n_2)\geq 0 \text{ and } \widetilde{f}_2(n_1,0)\geq 0.$$

*Proof of "Assumption 5.4 implies positivity".* We prove that if the initial data $n_1^{\text{init}}$, $n_2^{\text{init}}$ are non-negative and if Assumption 5.4 holds, then $n_1$ and $n_2$ remain non-negative. It is a simple application of the comparison principle for this parabolic system. A solution that lies initially above a sub-solution remains above it. The constant $(0,0)$ is indeed a sub-solution. □

For the sake of clarity of the presentation, we only consider an extension of the biological example from Section 5.3. This allows us to take into account imperfect maternal transmission. We assume that at a rate $\mu$, infected females lay eggs which do not carry *Wolbachia*. This quantity is very commonly tested by entomologists, and usually shown to be close to 0 (see [233] and references, and for example [75] where they obtained $\mu=0.04$ and $\mu=0$). This feature is included in the following model taken from [83] (neglecting the pathogen effect),

$$\begin{cases} \partial_t n_i - \nabla\cdot(A(x)\nabla n_i) = n_i F_u(1-s_f)(1-\mu) - n_i(d_i+\sigma(n_i+n_u)), \\ \partial_t n_u - \nabla\cdot(A(x)\nabla n_u) = n_u F_u(1-s_h\frac{n_i}{n_u+n_i}) + \mu F_u(1-s_f)n_i - n_u(d_u+\sigma(n_i+n_u)). \end{cases} \tag{5.33}$$

Here, the reduced population would be $n=\sigma(n_i+n_u)$. The corresponding dynamics in $(n,p)$ reads,

$$\begin{cases} \partial_t n - \nabla\cdot(A(x)\nabla n) = n\Big(F_u\big(p(1-s_f)+(1-p)(1-s_hp)\big) - d_u\big((\delta-1)p+1\big) - n\Big), \\ \partial_t p - \nabla\cdot(A(x)\nabla p) - 2\frac{\nabla n}{n}A(x)\nabla p = p\Big((1-p)\big(F_u(1-s_hp) - d_u(\delta-1)\big) - \mu F_u(1-s_f)\Big). \end{cases} \tag{5.34}$$

We notice in particular that the reaction term for $p$ in (5.34) does not depend on $n$. It yields directly the equation (5.3) with a function $n$ in the left hand side that depends on $p$, whereas in [29] the function $n$ in the gradient in the left hand side is assumed to be given.

As in Section 5.3, we introduce the parameter $\epsilon$ to characterize the high fertility and strong competition and propose the following extension of system (5.16), with imperfect maternal transmission,

$$\begin{cases} \partial_t n_i - \nabla\cdot(A(x)\nabla n_i) = (1-\mu)(1-s_f)F_u n_i\big(\frac{1}{\epsilon}-\sigma(n_i+n_u)\big)_+ - \delta d_u n_i, \\ \partial_t n_u - \nabla\cdot(A(x)\nabla n_u) = F_u\big(n_u(1-s_hp)+\mu(1-s_f)n_ip\big)\big(\frac{1}{\epsilon}-\sigma(n_i+n_u)\big)_+ - d_u n_u, \end{cases} \tag{5.35}$$

with $p=\frac{n_i}{n_i+n_u}$ as usual. In this system, the notation $a_+=\max\{0,a\}$ denotes the positive part of $a\in\mathbb{R}$.

For the reduction, as above, we identify $n_1=n_i$ and $n_2=n_u$ and we deduce from (5.35) the equations satisfied by $n=\frac{1}{\epsilon}-\sigma(n_i+n_u)$ and $p$,

$$\begin{aligned} \partial_t n - \nabla\cdot(A(x)\nabla n) = \; & -\big(\frac{1}{\epsilon}-n\big)F_u\left((1-s_f)((1-\mu)p+\mu p^2)+(1-p)(1-s_hp)\right)n_+ \\ & +d_u(\delta p+1-p)\big(\frac{1}{\epsilon}-n\big), \end{aligned} \tag{5.36}$$

$$\begin{aligned}\partial_t p - \nabla\cdot(A(x)\nabla p) + 2\frac{\nabla n}{n}A(x)\nabla p = \quad & F_u p\Big((1-p)\big((1-\mu)(1-s_f)-(1-s_h)p\big)\\ & +\mu(1-s_f)p^2\Big)n_+ + p(1-p)d_u(1-\delta).\end{aligned} \tag{5.37}$$

Using the notation in (5.5), we define as in (5.8) the function $H$ by

$$\begin{aligned}H(n,p) &:= -F_u n\big(p(1-\mu)(1-s_f)+(1-p)(1-s_h p)+\mu(1-s_f)p^2\big)+d_u(p(\delta-1)+1)\\ &= -F_u n\big((s_h+\mu(1-s_f))p^2-(s_f+s_h+\mu(1-s_f))p+1\big)+d_u((\delta-1)p+1).\end{aligned}$$

When $\mu = 0$, we notice that we recover the same expression as in the case of perfect maternal transmission in Section 5.3. Then, the function $h$ and the reaction term are modified.

In this case, as in Lemma 5.2, we may investigate the equilibria of (5.36)–(5.37). We get from straightforward computations:

**Lemma 5.7.** *Let*

$$\Delta = \big(\delta(s_f+s_h)+(\delta-1-\mu)(1-s_f)\big)^2 - 4\delta\big(s_h+\mu(1-s_f)\big)\big(\delta-(1-\mu)(1-s_f)\big).$$

*Let us assume that $\Delta > 0$. When $\mu = 0$, the condition $\Delta > 0$ is equivalent to $\big(\delta s_h-\delta+(1-s_f)\big)^2 > 0$ which is always satisfied. Then, there are 4 equilibria associated to the system (5.37)–(5.36) in the reduced variable $(n,p)$:*

- *The co-existence equilibrium reads*

$$\begin{cases} p_C^* = 1 - \dfrac{\delta(s_f+s_h)+(\delta-1+\mu)(1-s_f)-\sqrt{\Delta}}{2\delta(s_h+\mu(1-s_f))},\\ n_C^* = \dfrac{\delta d_u}{(1-\mu)(1-s_f)F_u}, \end{cases}$$

  *it remains unstable.*

- *The steady state $(0,0)$ is unstable.*

- *The stable Wolbachia invasion equilibrium reads*

$$\begin{cases} p_W^* = 1 - \dfrac{\delta(s_f+s_h)+(\delta-1+\mu)(1-s_f)+\sqrt{\Delta}}{2\delta(s_h+\mu(1-s_f))} < 1,\\ n_W^* = \dfrac{\delta d_u}{(1-\mu)(1-s_f)F_u} = n_C^*. \end{cases}$$

- *The stable Wolbachia extinction equilibrium is unchanged: $n_E^* = \dfrac{d_u}{F_u}$, $p_E^* = 0$.*

From straightforward computation, we may adapt Theorem 5.1 in this framework. Then, the analogue of Corollary 5.1 reads

**Corollary 5.2.** *Assume that $A$ satisfies Assumption 5.1. Given $n_1^{init,\epsilon}$ and $n_2^{init,\epsilon}$ such that there exists $p^{init} \in L^2(\mathbb{R}^d)$ such that $p^{init,\epsilon} \rightharpoonup p^{init}$ as $\epsilon \to 0$ in $L^2(\mathbb{R}^d)$-weak and $\frac{1}{\epsilon} - \sigma(n_1^{init,\epsilon}+n_2^{init,\epsilon}) - \frac{d_u}{F_u} \in L^2 \cap L^\infty(\mathbb{R}^d)$ with uniform bounds in $\epsilon > 0$, then Theorem 5.1 applies and the solutions $(n_i^\epsilon, n_u^\epsilon)_{\epsilon>0}$ of (5.35) satisfy the convergence result in (5.12). The limiting equation reads*

$$\partial_t p - \nabla\cdot(A(x)\nabla p) = r_\mu(p), \tag{5.38}$$

*where*

$$r_\mu(p) = d_u p\Big((1-\mu)(1-s_f)\frac{(\delta-1)p+1}{(s_h+\mu(1-s_f))p^2-(s_f+s_h+\mu(1-s_f))p+1} - \delta\Big).$$

For small $\mu$, $r_\mu$ is still a bistable function provided $\Delta > 0$, however the stable state 1 is displaced. The profile of this function appears on Figure 5.2

We give a numerical illustration of this case in Figure 5.3. We use the same parameters as for Figure 5.1, except that $s_f = 0$, $\mu = .04$ and the initial data is smaller (less infected mosquitoes are introduced).

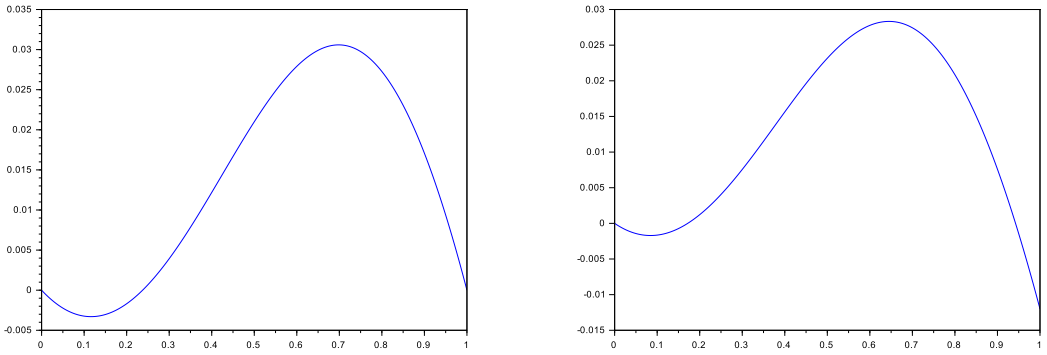


Figure 5.2: Bistable profiles of the reaction terms in (5.17) (left) and (5.38) (right), between 0 and 1.

For Figure 5.3, we use the same discretization and numerical scheme as in Figure 5.1. The blue lines represent the solution of the limiting system (5.38). In dashed red lines are plotted the numerical results for the system of two populations (5.37). We observe that the solution of the limiting bistable system (5.38) exhibits a traveling front which propagates into the whole domain. Then the numerical results for 4 different values of the parameter $\epsilon$ are represented. For large populations, we observe that as $\epsilon$ goes to 0 (recall that the order of magnitude of the population size is $\frac{1}{\sigma\epsilon}$), the solution to the whole system (5.37) gets closer to the one of the limiting system. However, for small populations, we see a clear modification of the wave's shape and speed, which is slower than the limit wave.

## 5.6 Conclusion and perspectives

We have established in this paper the rigorous convergence, under suitable assumptions, of a 2 by 2 reaction diffusion model of Lotka-Volterra type towards a simple model for the frequency of a variant. It justifies the use of such reduced model in applications. Let us discuss quickly our scaling choice in Assumption 5.2, in the case of *Wolbachia*.

Another biologically relevant scaling assumption would not give a limiting system consisting in only one equation on frequency. Indeed, if we consider the following alternative model

$$\begin{cases} \partial_t n_i - \nabla\cdot(A(x)\nabla n_i) &= (1-s_f)F_u n_i\big(1-\epsilon\sigma(n_i+n_u)\big) - \delta d_u n_i, \\ \partial_t n_u - \nabla\cdot(A(x)\nabla n_u) &= F_u n_u(1-s_h\frac{n_i}{n_i+n_u})\big(1-\epsilon\sigma(n_i+n_u)\big) - d_u n_u. \end{cases} \tag{5.39}$$

Then, $n$ and $p$ satisfy the following system, that does not depend on $\epsilon$

$$\begin{cases} \partial_t n - \nabla\cdot(A(x)\nabla n) = F_u\big(1-n\big)\big(A(p)-B(p)n\big), \\ \partial_t p - \nabla\cdot(A(x)\nabla p) + \frac{2\nabla p\cdot A(x)\nabla n}{1-n} = p(1-p)\big(F_u n(s_h p - s_f) - d_u(\delta-1)\big), \end{cases} \tag{5.40}$$

where

$$\begin{cases} A(p) &= ((\delta-1)p+1)\frac{d_u}{F_u}, \\ B(p) &= s_h p^2 - (s_f+s_h)p + 1. \end{cases}$$

The dependency in $\epsilon$ in the resulting model is only through the initial data. Thus, $\epsilon\to 0$ does not imply $n - \frac{A(p)}{B(p)} \to 0$ in (5.39), (5.40).

We conclude that the use of simple bistable models for the spatial spread of *Wolbachia* can be justified mathematically. This is the object of Theorem 5.1. However, we must keep in mind that this result applies only if population size and fecundity scale properly.

In the context of *Wolbachia* modeling, bistable equations like (5.3) have been used (for example in [29] or [203]) because they provide with a unique (up to translations) and linearly stable traveling wave solution. Hence, with a bistable model at hand we can compute a speed that may be interpreted as an invasion speed.

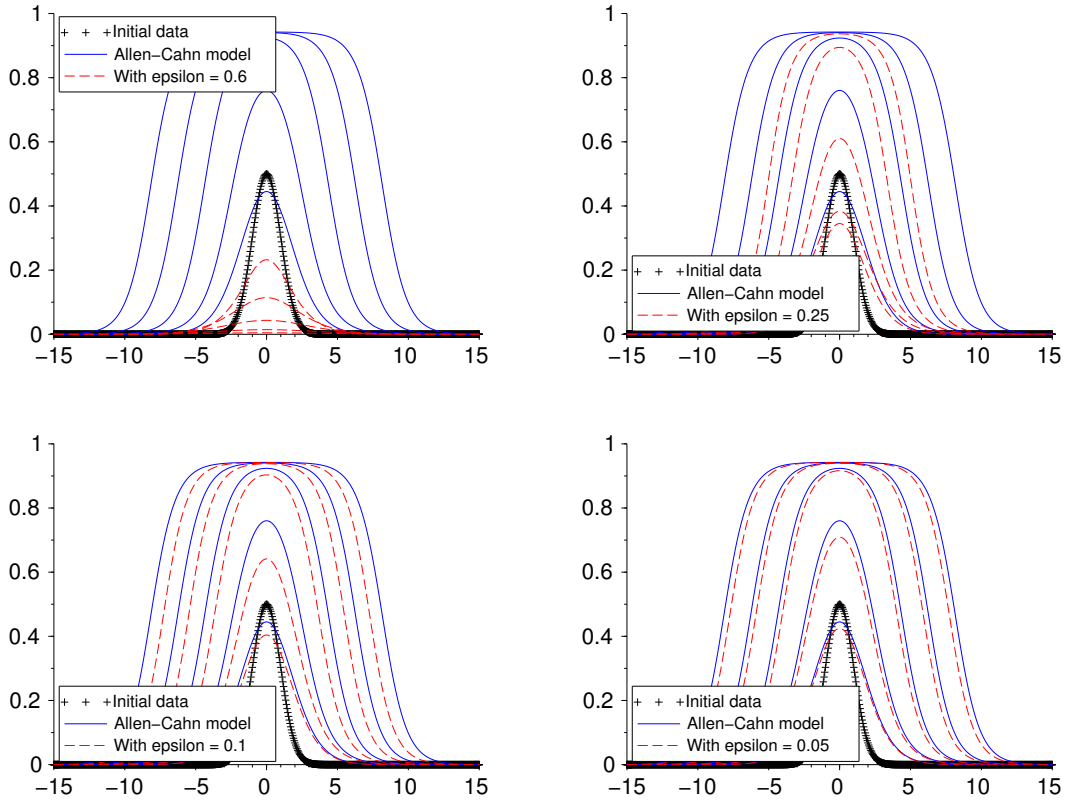



Figure 5.3: Initial data (+) creating a traveling wave in the limit system (blue) and convergence of the two-population solution (dashed red) as $\epsilon$ diminishes.

Therefore a natural continuation of the present work would be to try and specify Theorem 5.1 to traveling waves. The open question reads: does the frequency in the two-populations model converge to the unique traveling wave solution of the limit bistable equation? If yes, in what sense? Indeed, there are two types of convergence involved: on the first hand in the singular limit (where we identified a small parameter $\epsilon$), that proves convergence of the system's frequency to a solution of the limit bistable equation; and on the other hand the well-known attractiveness result of the unique traveling wave solution in the bistable case. Moreover, existence and (local) stability of traveling waves has been proved for competitive systems (see [91] for example). How to compare the traveling speed for competitive system with the one for the reduced model on the frequency?

**Acknowledgements.** The authors acknowledge partial supports from the Capes/Cofecub project Ma-833 15 *"Modeling innovative control method for Dengue fever"* and from the Programme Convergence Sorbonne Universités / FAPERJ *"Control and identification for mathematical models of Dengue epidemics"*.

They warmly thank B. Perthame for his patient and constant help, useful discussions and valuable suggestions on the manuscript.

# Chapter 6

# Hindrances to bistable propagation: wave-blocking, wave-delaying

> Si quelque chose s'oppose à toi et te déchire, laisse croître, c'est que tu prends racine et que tu mues.
>
> Antoine de Saint-Exupéry, *Citadelle*.

This chapter is a joint work with Nicolas Vauchelet and Grégoire Nadin. It was published as an article in the Journal of Mathematical Biology [176]. Compared with the published version, the reminder section 6.3 has been moved to the context presentation of the thesis (Section 4.3.2) and an appendix has been added to give some additional results. Therefore, the references to this manuscript that appear in the published version have been redirected to Appendix 6.A. In addition, four misprints showing $X_\alpha^{-1}$ instead of $X_\alpha$ have been corrected, and also a missing 2 in the expression of $L_*(C)$ given in the proof of Proposition 6.13.

**Abstract.** We study the biological situation when an invading population propagates and replaces an existing population with different characteristics. For instance, this may occur in the presence of a vertically transmitted infection causing a cytoplasmic effect similar to the Allee effect (*e.g.* *Wolbachia* in *Aedes* mosquitoes): the invading dynamics we model is bistable.

We aim at quantifying the propagules (what does it take for an invasion to start?) and the invasive power (how far can an invading front go, and what can stop it?).

We rigorously show that a heterogeneous environment inducing a strong enough population gradient can stop an invading front, which will converge in this case to a stable front. We characterize the critical population jump, and also prove the existence of unstable fronts above the stable (blocking) fronts. Being above the maximal unstable front enables an invading front to clear the obstacle and propagate further.

We are particularly interested in the case of artificial *Wolbachia* infection, used as a tool to fight arboviruses.

## 6.1 Introduction

The fight against world-wide plague of dengue (see [32]) and of other arboviruses has motivated extensive work among the scientific community. Investigation of innovative vector-control techniques has become a well-developed area of research. Among them, the use of *Wolbachia* in *Aedes* mosquitoes to control diseases (see [232, 8]) has received considerable attention. This endosymbiotic bacterium is transmitted from mother to offspring, it induces cytoplasmic incompatibility (crossings between infected males and uninfected females are unfertile) and blocks virus replication in the mosquito's body. Artificial infection can be performed in the lab, and vertical transmission (from mother to progeny) allows quick and massive rearing of an infected colony. Pioneer mathematical modeling works on this technique include [29, 121, 102].

We are mostly interested in the way space interferes during the vector-control processes. More precisely, we would like to understand when mathematical models including space can effectively predict the blocking of an on-going biological invasion, which may have been caused, for example, by releases during a vector-control program.

The observation of biological invasions, and of their blocking, has a long and rich history. We simply give an example connected with *Wolbachia*. In the experimental work [15], it was proved that a stable coexistence of several (three) natural strains of *Wolbachia* can exist, in a *Culex pipiens* population. The authors mentioned several hypotheses to explain this stability. Our findings in the present paper - using a very simplified mathematical model - partly supports the analysis conducted in the cited article. Namely, "differential adaptation" cannot explain the blocking, while a large enough "population gradient" can, and we are able to quantify the strength of this gradient, potentially helping validating or discarding this hypothesis.

Although the field experiments have not yet been conducted for a significantly long period, artificial releases of *Wolbachia*-infected mosquitoes (see [118, 178]) also seem to experience such "stable fronts" or blocking phenomena (see [117, 239]). This issue was studied from a modeling point of view in [29, 50] (reaction-diffusion models), [100] (heterogeneity in the habitat) and [103] (density-dependent effects slowing the invasion), among others.

In order to represent a biological invasion in mathematical terms as simply as possible, reaction-diffusion equations have been introduced (for the first time in [87] and [138]) in the form

$$\partial_t u - \Delta u = f(u), \tag{6.1}$$

where $t \geq 0$ and $x \in \mathbb{R}^d$ are respectively time and space variables, $d$ is the spatial dimension and $u(t,x)$ is a density of alleles in a population, at time $t$ and location $x$. This very common model to study propagation across space in population dynamics enhances a celebrated and useful feature: existence (under some assumptions on $f$) of traveling wave solutions. In space dimension 1, a traveling wave is a solution $u(t,x) = \widetilde{u}(x - ct)$ to (6.1), where $c \in \mathbb{R}$, $\widetilde{u}$ is a monotone function from $\mathbb{R}$ to $\mathbb{R}$, and $\widetilde{u}(\pm\infty) \in f^{-1}(0)$. By convention, we will always use decreasing traveling waves. They have a constant shape and move at the constant speed $c$.

The quantity $u$ may represent the frequency of a given trait (phenotype, genotype, behavior, infection, etc.) in a population. In this case, the model below has been introduced in order to account for the effect of spatial variations in the total population density $N$ (see [26, 29]) in the dynamics of a frequency $p$

$$\partial_t p - \Delta p - 2\frac{\nabla N \cdot \nabla p}{N} = f(p). \tag{6.2}$$

The additional $\nabla(\log(N))$ term is known as the "gene flow". It represents the fact that the genotype of individuals from high-abundance areas tends to be over-represented (due to their number) in the offspring from neighboring low-abundance areas, due to dispersion (see [26]).

The effect of gene flow at the edge of the invasion fronts (where total population $N \simeq 0$) has been studied by many authors in numerous models, *e.g.* in [135], and with an adaptive point of view in [46] and [171]. They showed that it can pin an invasion and limit the range of an invading species. In the present article, although the total population $N$ stays uniformly positive, we still observe invasion pinning phenomena.

In [29] Barton and Turelli pointed out the fact that cytoplasmic incompatibility can be seen, from a modeling point of view, as an analogue of an Allee effect. Mathematically speaking, the main consequence is that both states $p = 0$ (the trait disappears) and $p = 1$ (the trait is everywhere) are locally stable, that is, we consider a bistable nonlinearity. Allee effects alone were proved numerically in [132] to be able to limit the range of an invading species, in discrete space models. For continuous models, it is known that one needs heterogeneous coefficients in order to observe a similar phenomenon (see [132]). Here we consider a continuous heterogeneous model where heterogeneity comes from the gene flow.

Indeed, we study a slightly different situation where the species has already sprawled all over the domain. In this context, gene flow does not affect the species' range but the spread of a new "trait" appearing in the population, by hindering trait propagation from low-abundance to high-abundance areas.

In some cases, the total population density $N$ may be affected by the trait frequency $p$, and even depend explicitly on it. In the large fecundity asymptotic for the spread of *Wolbachia* (we refer to a detailed derivation in Section 6.6.2), where $p$ stands for the infection frequency, it was

proved (in [211]) that there exists a function $h : [0,1] \to (0,+\infty)$ such that $N = h(p) + o(1)$ in the limit when fecundity goes to $+\infty$.

Hence we can write the first-order approximation

$$\partial_t p - \Delta p - 2\frac{h'(p)}{h(p)}|\nabla p|^2 = f(p). \tag{6.3}$$

Our main results are the characterization of the asymptotic behavior of $p$ in two settings: for equation (6.2) when $N$ only depends on $x$ (which can also be seen as a limit obtained from two-populations models, see the derivation in Section 6.6.2), and for equation (6.3) in all generality. Both of them may be seen as special cases of the general problem

$$\partial_t p - \Delta p - 2\nabla\big(V\big(x, p(t,x)\big)\big) \cdot \nabla p = f(p).$$

For (6.2) with $d = 1$, our characterization is sharp when $\partial_x \log N$ is equal to a constant times the characteristic function of an interval. Overall, two possible sets of asymptotic behaviors appear. On the first hand, the equation can exhibit a sharp threshold property, dividing the initial data between those leading to invasion of the infection ($p \to 1$) and those leading to extinction ($p \to 0$) as time goes to infinity. In this case, the threshold is constituted by initial data leading convergence to a ground state (positive non-constant stationary solution, going to 0 at infinity). It is a sharp threshold, which implies that the ground state is unstable. We show that such a threshold property always holds for equation (6.3), and occurs in some cases for equation (6.2). On the other hand, the infection propagation can be blocked by what we call here a "barrier" that is a stationary solution or, in the biological context, a blocked propagation front. We show that this happens in (6.2), essentially when $\partial_x \log N$ is large enough. This asymptotic behavior differs from convergence towards a ground state in the homogeneous case. Indeed, even though the solution converges towards a positive stationary solution, we prove that in this barrier case, the blocking is actually stable. Some crucial implications for practical purposes (use of *Wolbachia* in the field) of this stable failure of infection propagation are discussed.

Even though all our results are new, the pinning effect of gene flow in bistable models was already identified. In the context of Allee effect-induced bistability, [132] shows indeed that discrete models can predict invasion failure (or "pinning") while continuous models fail to do so in homogeneous environments. Our results confirm this intuition, and also depict an intermediate modeling level consisting of a continuous model with heterogeneous environment. We can give a precise characterization of the constant population gradient in a bounded area which is required to pin the invasion.

From the mathematical point of view, our work on (6.2) makes use of a phase-plane method that can be found in [148] (and also in [51] and [191]) to study similar problems. It helps getting a good intuition of the results, coupled with a double-shooting argument. We note that a shooting method was also used in [161] for ignition-type nonlinearity, in a non-autonomous setting, to get similar results under monotonicity assumptions we do not require here.

The paper is organized as follows. Main results on both (6.3) and (6.2) are stated in Section 6.2, where their biological meaning is explained. We also give illustrative numerical simulations. After a brief recall of well-known facts on bistable reaction-diffusion in Section 6.3, we prove our results on (6.3) in Section 6.4, and on (6.2) in Section 6.5. Finally, Section 6.6 is devoted to a discussion on our results, and on possible extensions. Moreover, because it was the work that first attracted us to this topic, we expand in Section 6.6.3 on the concept of local barrier developed by Barton and Turelli in [29], and relate it to the present article.

## 6.2 Main results

### 6.2.1 Statement of the results

#### Results on the infection-dependent case

Our first set of results is concerned with (6.3), where the total population is a function of the infection frequency.

We notice that the problem (6.3) is left invariant by multiplying $h$ by any $\lambda \in \mathbb{R}^*$. Without loss of generality we therefore fix $\int_0^1 h^2(\xi)\, d\xi = 1$, and state

**Theorem 6.1.** *Let $H$ be the antiderivative of $h^2$ which vanishes at 0, that is $H(x) := \int_0^x h^2(\xi)\,d\xi$. $H$ is a $\mathcal{C}^1$ diffeomorphism from $[0,1]$ into $[0,1]$.*

*Let $g : [0,1] \to [0,1]$ such that for all $x \in [0,1]$, $g(H(x)) = f(x)h^2(x)$.*

*There exists a traveling wave for (6.3) if and only if there exists a traveling wave for (6.1) with reaction term $g$ (i.e. for the equation $\partial_t u - \partial_{xx} u = g(u)$). In addition:*

1. *If $f$ satisfies the KPP (named after [138]) condition $f(x) \leq f'(0)x$ and if $H$ is concave (which is equivalent to $h' \leq 0$), then there exists a minimal wave speed $c_* := 2\sqrt{g'(0)}$ for traveling wave solutions to (6.3). This means that for all $c \geq c_*$, there exists a unique traveling wave solution to (6.3) with speed $c$.*

2. *If $f$ is bistable then there exists a unique traveling wave for (6.3). Its speed has the sign of*

$$\int_0^1 f(x)h^4(x)dx.$$

Depending on the initial data, in this case, solution can converge to 1 ("invasion"), initiating a traveling wave with positive speed, or to 0 ("extinction"). Note that non-constant $h$ may have a huge impact in the asymptotic behavior, possibly reversing the traveling wave speed: in this case, 0 would become the invading state instead of 1.

In the case of *Wolbachia*, we discuss the expression of $h$ in Subsection 6.4.1, and give a numerical example of this situation in Subsection 6.2.3.

We can construct a family of compactly supported "propagules", that is functions which ensure invasion.

**Proposition 6.1.** *There exists $\theta_c \in (0,1)$ (defined below by (4.11)) such that for all $\alpha \in (\theta_c, 1)$, there exists $v_\alpha \in \mathcal{C}^2_p(\mathbb{R}, [0,\alpha])$ ($v_\alpha$ is continuous and piecewise twice continuously differentiable), whose support is equal to $[-L_\alpha, L_\alpha]$ for a known $L_\alpha \in (0, +\infty)$ (given below by (6.12)), such that $0 \leq v_\alpha \leq \alpha$, $\max v_\alpha = v_\alpha(0) = \alpha$, $v_\alpha$ is symmetric and radial-non-increasing, and $v_\alpha$ is a sub-solution to (6.3).*

We call $v_\alpha$ an $\alpha$-bubble (associated with (6.3)), or $\alpha$-propagule, following the definition in [29].

### Results on the heterogeneous case

Our second set of results deals with the situation where the total population of mosquitoes strongly increases in a given region of the domain. In this case, the total population $N$ is given and we consider the model (6.2). Before stating our main result on equation (6.2), we introduce the concept of *propagation barrier* (which we will simply call *barrier* below).

To fix the ideas and get a tractable problem, we assume that $N$ increases (exponentially) in a given region of spatial domain and is constant in the rest of the domain. We consider that the domain is one-dimensional and therefore investigate the differential equation

$$\partial_t p - \partial_{xx} p - 2\partial_x(\log N)\partial_x p = f(p). \tag{6.4}$$

In view of the setting we have in mind for $N$ we let, for some $C, L > 0$:

$$\partial_x \log(N) = \begin{cases} \dfrac{C}{2}, & \text{on } [-L, L], \\ 0, & \text{on } \mathbb{R} \setminus [-L, L]. \end{cases} \tag{6.5}$$

Existence of a stationary wave for this problem boils down to the existence of a solution to

$$\begin{cases} -p'' - Cp' = f(p), & \text{on } [-L, L], \\ -p'' = f(p), & \text{on } \mathbb{R} \setminus [-L, L], \\ p(-\infty) = 1, \quad p(+\infty) = 0, \quad p > 0, \end{cases} \tag{6.6}$$

which is a well-defined problem in the space of continuously differentiable real functions which are twice continuously differentiable on $(-L, L)$ and on $\mathbb{R}\setminus[-L, L]$.

In the context of our study, stationary solutions to (6.4) with prescribed behavior at infinity, that is solutions of (6.6), play the role of *barriers*, blocking the propagation of the infection.

**Definition 6.1.** *We name a $(C, L)$-barrier any solution to (6.6). For any bistable function $f$ we define the barrier set*

$$\mathcal{B}(f) := \big\{(C, L) \in (0, +\infty)^2, \text{ there exists a } (C, L)\text{-barrier}\big\}. \tag{6.7}$$

As we will recall in Section 6.3, in the bistable case there exists a unique (up to translations) traveling wave solution to (6.1). This solution can be seen as a solution to the limit problem of (6.6) as $L \to +\infty$. We make this intuition more precise in this paper (see in particular Proposition 6.4 below).

The bistable traveling wave is associated with a unique speed that we denote $c_*(f)$ (see Section 6.3 for definitions and a brief review of classical results on bistable reaction-diffusion equation).

**Theorem 6.2.** *Let $C > 0, L > 0$ and assume $N$ is given by (6.5). For $C > c_*(f)$, there exists $L_*(C) \in (0, +\infty)$ such that $(C, L) \in \mathcal{B}(f)$ if and only if $L \geq L_*(C)$.*

Existence of a barrier, as stated in Theorem 6.2, has strong and direct consequences on the asymptotic behavior of solutions to (6.2).

**Proposition 6.2.** *Assume $N$ is defined by (6.5). If $(C, L) \in \mathcal{B}(f)$ we denote by $p_B$ a solution to the standing wave problem (6.6). Then any solution of (6.4) with initial value $p^0$ satisfying $p^0 \leq p_B$ has limited propagation, which means that $\forall x \in \mathbb{R}, \limsup_{t\to\infty} p(t, x) < 1$. More precisely,*

$$\forall t \geq 0, \quad p(t, x) \leq p_B(x).$$

*On the contrary, assume that either (6.6) has no solution (i.e. $(C, L) \notin \mathcal{B}(f)$) and $p^0$ has a limit at $-\infty$ equal to 1, or there exists a solution $p_B$ to (6.6) which is unstable from above (in the sense of Definition 4.5), such that $p_0 > p_B$ and there is no other solution $p_{B'}$ to (6.6) satisfying $p_{B'} > p_B$. In this case $p$ propagates, that is:*

$$\forall x \in \mathbb{R}, \quad \limsup_{t\to\infty} p(t, x) = 1.$$

We also characterize the barriers

**Proposition 6.3.** *Let $(C, L) \in \mathcal{B}(f)$. Then*

1. *Any $(C, L)$-barrier (i.e. solution of (6.6)) is decreasing.*
2. *If $L > L_*(C)$ then there exists at least two $(C, L)$-barriers.*
3. *The $(C, L)$-barriers are totally ordered (in the sense that given two $(C, L)$-barriers $p_B$ and $p_{B'}$ then either $p_B(x) \leq p_{B'}(x)$ for all $x \in \mathbb{R}$ or $p_B(x) \geq p_{B'}(x)$ for all $x \in \mathbb{R}$), hence we can define a maximal and a minimal element among them.*
4. *The maximal $(C, L)$-barrier is unstable from above and the minimal one is stable from below (in the sense of Definition 4.5 below).*

We also get a picture of the behavior of $L_*(C)$:

**Proposition 6.4.** *The function $L_*$ is decreasing and satisfies*

$$\lim_{C\to c_*(f)^+} L_*(C) = +\infty, \quad L_*(C) \sim \frac{1}{4C} \log\big(1 - \frac{F(1)}{F(\theta)}\big) \text{ when } C \to +\infty.$$

Instead of restricting to a constant (logarithmic) population gradient, we can very well let it vary freely. To do so we introduce a set of gradient profiles which we denote by $\mathcal{X}$. For example,

$$\mathcal{X} := \{h : \mathbb{R} \to \mathbb{R}_+, \, h \in L^\infty \text{ with compact support.}\} \tag{6.8}$$

Then, the barriers may be defined in a similar fashion as before.

**Definition 6.2.** *For $h \in \mathcal{X}$, a h-barrier is any solution to the "standing wave equation"*

$$\begin{cases} -p'' - h(x)p' = f(p) \text{ on } \mathbb{R}, \\ p(-\infty) = 1, \quad p(+\infty) = 0. \end{cases} \tag{6.9}$$

*We define the barrier set associated with (6.8)*

$$\mathcal{B}_{\mathcal{X}}(f) := \{h \in \mathcal{X}, \text{ there exists a } h\text{-barrier}\}.$$

In this setting, a meaningful extension of Theorem 6.2 is the following

**Corollary 6.1.** *Let $h \in \mathcal{X}$. If $(C, L) \in \mathcal{B}(f)$ and $h \geq C\mathbb{1}_{[-L,L]}$ then $h \in \mathcal{B}_{\mathcal{X}}(f)$. Conversely, if $(C, L) \notin \mathcal{B}(f)$ and $h \leq C\mathbb{1}_{[-L,L]}$ then $h \notin \mathcal{B}_{\mathcal{X}}(f)$.*

**Remark 6.1.** *We do not require more regularity on functions in $\mathcal{X}$ because Corollary 6.1 only relies on a comparison principle.*

### 6.2.2 Biological interpretation

Our results on possible propagation failures can be summarized and interpreted easily.

On the first hand, if the size of the population is regulated only by the level of the infection (or the trait frequency), then in a homogeneous medium no stable blocked front can appear (this is the sharp threshold property implied by Theorem 6.1), except in the very particular case when $\int_0^1 f(x)h^4(x)dx = 0$. This situation can be understood as the limit when local demographic equilibrium is reached much faster than the infection process (or when the population is typically large, as in the asymptotic from [211]), which makes sense in the context of *Wolbachia* because the infection is vertically transmitted.

On the second hand, if the carrying capacity (or "nominal population size") is heterogeneous (in space), then an increase in the population size raises a hindrance to propagation, that can be sufficient to effectively block an invading front (Theorem 6.2), and give rise to a stable transition area (as observed in [15]), even if the infection status does not modify the individuals' fitness. This situation is particularly adapted to a wide range of *Wolbachia* infections, when several natural or artificial strains do not have very different impacts on the host's fitness. We note that the case when the heterogeneity concerns the diffusivity rather than the population size was treated in [148], yielding the same conclusion: a large-enough area of low-enough diffusivity stops the propagation.

From our results, we draw two conclusions that are relevant in the context of biological invasions.

First, fitness cost (and cytoplasmic incompatibility level, in the case of *Wolbachia*) determines the existence of an invading front in a homogeneous setting, and eventually its speed. However, ecological heterogeneity (rather than fitness cost) seems to play a prominent role in propagation failure - or success - of a given infection.

Second, the existence of a stable (from below) front implies the existence of an unstable (from above) one, as stated in Proposition 6.3. Therefore, any of the heterogeneity-induced hindrances to propagation that have been identified (here and in [148]) can be jumped upon. It suffices that the infection wave reaches the unstable front level. Computing the location and level of this theoretical "unstable front", in the presence of an actual "stable front", is extremely useful: either to estimate the risk that the infection propagates through the barrier into the sound area, or to know the cost of the supplementary introduction to be performed in order to propagate the infection through the obstacle (in the case of blocked propagation following artificial releases of *Wolbachia*, for example, as seems to be the case in the experimental situation described in [117]).

### 6.2.3 Numerical illustration

Figure 6.1 is an illustration of Theorem 6.1. We choose $f$ and $h$ from the case of *Wolbachia* (see discussion on $h$ in Subsection 6.4.1) with perfect vertical transmission and biological parameters selected after the choices in [211]:

$$f(p) = d_s p \frac{-s_h\delta p^2 + \big(\delta(1+s_h) - (1-s_f)\big)p + (1-s_f) - \delta}{s_h p^2 - (s_f + s_h)p + 1},$$
$$h_\epsilon(p) = 1 - \epsilon\frac{d_u}{\sigma F_u}\frac{(\delta-1)p+1}{s_h p^2 - (s_f+s_h)p+1},$$

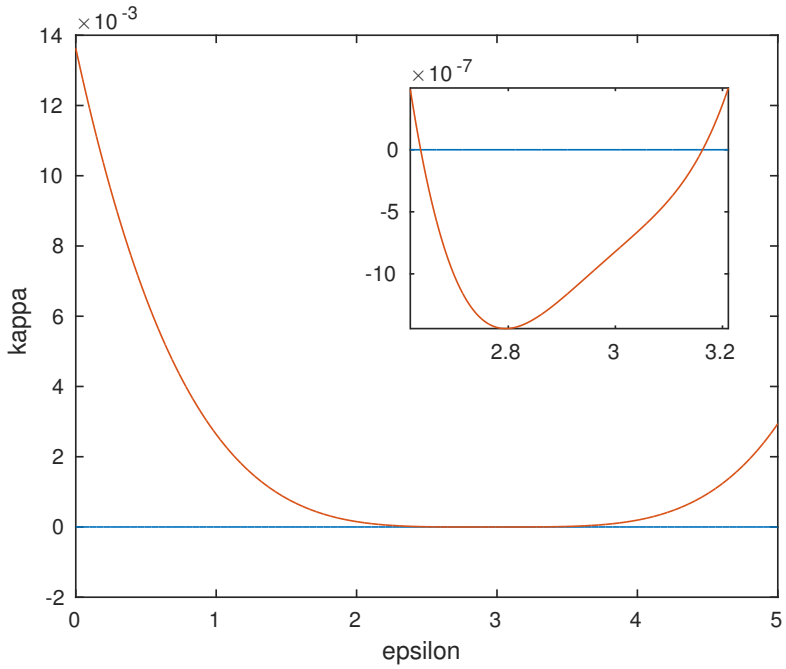



Figure 6.1: The function $\kappa : \epsilon \mapsto \int_0^1 f(p)h_\epsilon^4(p)dp$, whose sign is equal to that of the bistable traveling wave speed. The top-right angle plot is a zoom in the region where this sign is negative.

and $\kappa(\epsilon) = \int_0^1 f(p)h_\epsilon(p)dp$. We stick to this choice of $f$ for the other figures of this paper.

Figures 6.2, 6.3 and 6.4 must be interpreted as follows: the $y$-axis, oriented to the bottom, is time $t \in [0, 400]$, while the $x$-axis is the space, $x \in [-20, 20]$. The value of $p(t, x) \in [0, 1]$ is represented by a color, with the legend on the right-side of the plots. Simulations were done using a centered finite-difference scheme for diffusion and Euler implicit for time, with discretization steps $\Delta t = 0.05$ in time and $\Delta x = 0.1$ in space. Vertical dotted red lines mark the spatial range (=support) of the population gradient.

Figures 6.2 and 6.3 are illustrations of Proposition 6.2. On Figure 6.2, the two plots differ only by the value of the population gradient $C$ (respectively equal to 2 and 1), imposed in both cases on the interval $[-0.5, 0.5]$. The initial data is front-like, *i.e.* equal to 1 on $[-20, -14]$. On Figure 6.3, the population gradient is fixed at $C = 0.35$ with $L = 3$. The two plots differ by their initial data: they are still front-like, but on $[-20, -15]$ on the left-hand side, and on $[-20, 2]$ on the right-hand side. On Figure 6.2, on the left-hand plot we notice that a wave forms and propagates at a constant speed before being blocked, giving rise to a stable front ; while on the right-hand plot, the propagation occurs, and its speed is perturbed first by the heterogeneity, and then by the boundary of the discretization domain. The interpretation is similar for Figure 6.3.

Then, Figure 6.4 is an illustration of Corollary 6.1: it reproduces the behavior shown in Figure 6.2 for more sophisticated population gradients. We choose $h(x) = 4C(x - L)(x + L)/L^2$, with $L = 6$ and respectively $C = 0.5$ (left-hand side) and $C = 0.2$ (right-hand side), yielding blocking or propagation.

Finally Figures 6.5 and 6.6 illustrate Proposition 6.4. Because of the high convergence speed of $CL_*(C)$ towards its finite limit for large $C$, we draw its logarithm in Figure 6.6 to get a better picture of convergence order.

We also note on Figure 6.6 that $C \mapsto CL_*(C)$ appears to be decreasing. We were only able to prove this fact asymptotically (as $C \to \infty$) and we refer to Appendix 6.A for the explicit computations.

## 6.3 A brief reminder on bistable reaction-diffusion in $\mathbb{R}$

We refer to Section 4.3.2.

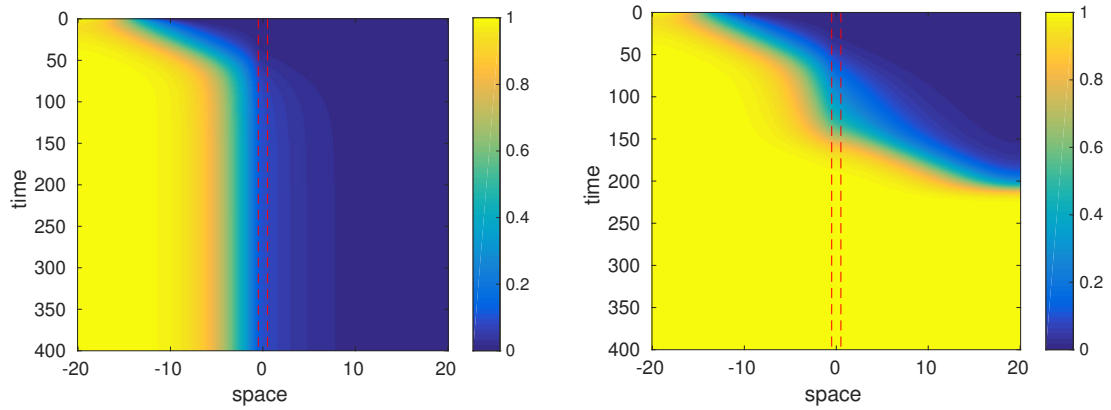



Figure 6.2: Plot of the proportion $p$ of the invading population in the total population with respect to time (y-axis) and space (x-axis). Two different population gradients are used with the same front-like initial data. The vertical red dotted lines mark the region $[-L, L]$ where the spatial gradient is applied. *Left*: Blocking with $L = 0.5$ and $C = 2$. *Right*: Propagation with $L = 0.5$ and $C = 1$.

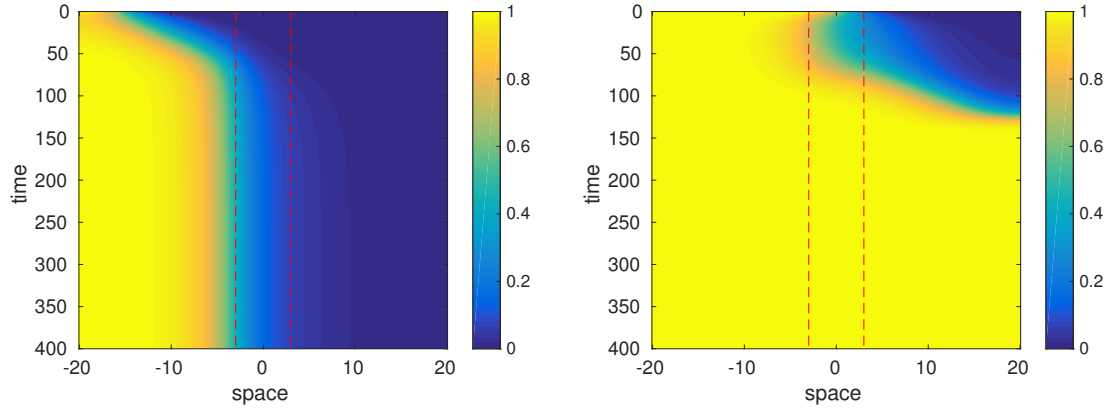



Figure 6.3: Plot of the proportion $p$ of the invading population in the total population with respect to time (y-axis) and space (x-axis). Two different front-like initial data are used with the same population gradient, $L = 3$ and $C = 0.35$. The vertical red dotted lines mark the region $[-L, L]$ where the spatial gradient is applied. *Left*: Blocking with a Heaviside initial datum located at $-15$. *Right*: Propagation with a Heaviside initial datum located at 2.

## 6.4 Proofs for the infection-dependent population gradient model

We recall equation (6.3), in dimension $d = 1$, for which we are going to prove Theorem 6.1

$$\partial_t p - \partial_{xx} p - 2\frac{h'(p)}{h(p)}|\partial_x p|^2 = f(p).$$

After giving an expression for $h$ in the case of *Wolbachia*, we prove that there exist traveling wave solutions to (6.3), whose speed sign can be determined easily, and eventually compared with traveling waves for (6.1). They can be initiated by "$\alpha$-propagules" (or "$\alpha$-bubbles") as in the case of (6.1), which was studied in [29] and [212]. Due to the classical sharp-threshold phenomenon for bistable reaction-diffusion (see [243] for the first proof with initial data as characteristic functions of intervals, [190] for extension to higher dimensions, [70, 163] and [174] for extension to localized initial data in dimension 1) solutions then have a simple asymptotic behavior. The infection can either invade the whole space or extinct (or, for a "lean" set of initial data, converge to a ground state profile, and this is an unstable phenomenon).

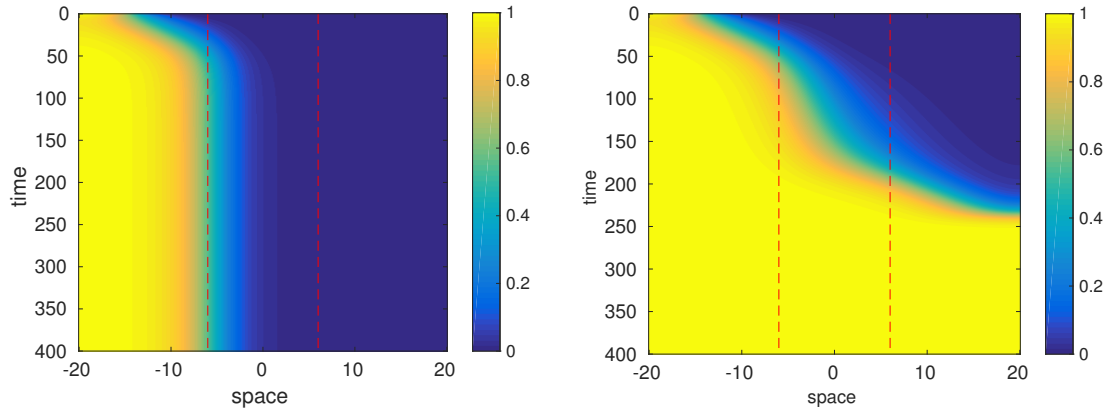



Figure 6.4: Plot of the proportion $p$ of the invading population in the total population with respect to time (y-axis) and space (x-axis). Two different, nontrivial population gradients ($h(x) = 4C(x-L)(x+L)/L^2$) are used, with the same front-like initial data. The vertical red dotted lines mark the region $[-L, L]$ where the spatial gradient is applied. *Left*: Blocking with $L = 6$, $C = 0.5$. *Right*: Propagation with $L = 6$, $C = 0.2$.

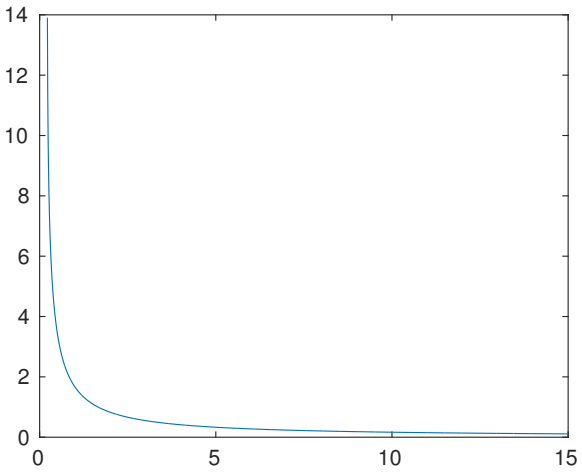



Figure 6.5: The minimal interval length $C \mapsto L_*(C)$ for which a logarithmic gradient constant equal to $C$ is sufficient to block invasion.

Hence when the population gradient is a function of the infection rate, there is no wave-blocking phenomenon.

### 6.4.1 In the case of *Wolbachia*, $h$ is not monotone

Clearly, if $h$ is non-increasing, $h' \leq 0$, then the solution $p$ to (6.3) is a sub-solution to (6.1), assuming we complete them with the same initial data. Hence $p \leq u$ for all time.

However, in the case of *Wolbachia*, the function $h$ (computed in the large population asymptotic developed in [211]) is not monotone. It reads

$$N = h(p) = 1 - \epsilon \frac{d_u}{\sigma F_u} \frac{(\delta - 1)p + 1}{s_h p^2 - (s_f + s_h)p + 1},$$

hence

$$h'(p) = \epsilon \frac{d_u}{\sigma F_u} \frac{(\delta - 1)s_h p^2 + 2s_h p - (\delta - 1 + s_f + s_h)}{\left(s_h p^2 - (s_f + s_h)p + 1\right)^2}.$$

We can compute $h'(0) < 0$, $h'(1) > 0$, for $\delta s_h - \delta + 1 - s_f > 0$ (this condition being necessary to ensure bistability in the limit equation, see details in [211]).

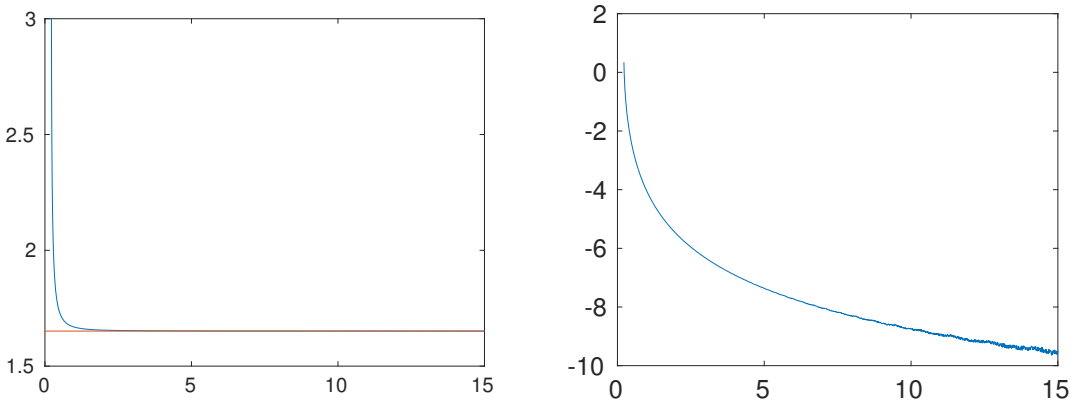



Figure 6.6: *Left*: The curve $C \mapsto 4CL_*(C)$ converges to the constant $\log\big(1 - F(1)/F(\theta)\big)$. *Right*: Visualization of the exponential rate of convergence: $C \mapsto \log\Big(4CL_*(C) - \log\big(1 - \frac{F(1)}{F(\theta)}\big)\Big)$.

We can show that $h'$ vanishes at a single point in $[0,1]$, where its sign changes. This point is

$$\theta_0 := \frac{1}{\delta - 1}\left(-1 + \sqrt{1 + (\delta - 1)(\frac{\delta - 1 + s_f}{s_h} + 1)}\right)$$

for $\delta \neq 1$, and if $\delta = 1$, then $\theta_0 = \dfrac{1}{2} + \dfrac{s_f}{2s_h}$.

Hence if $p \leq \theta_0$ then $h'(p) \leq 0$. As a consequence, for an initial datum $u^{\text{init}} = p^{\text{init}}$ such that $\|p^{\text{init}}\|_\infty \leq \theta_0$, $p \leq u$ holds as long as $\|p\|_\infty \leq \theta_0$. But no more can be said simply from (6.1).

### 6.4.2 A change of variable to recover traveling waves

*Theorem 6.1.* First, we note that the function $H(x) = \int_0^x h^2(\xi)d\xi$ is invertible on $[0,1]$, since it is increasing ($h^2 > 0$).

Multiplying (6.3) by $h^2(p)$ yields

$$h^2(p)\partial_t p - \partial_x(h^2(p)\partial_x p) = f(p)h^2(p).$$

We set $y(x) = H(p(x))$ (equivalently, $p(x) = H^{-1}(y(x))$). Then

$$\partial_t y - \partial_{xx} y = f(H^{-1}(y))h^2(H^{-1}(y)).$$

And we are left with the following problem

$$\partial_t y - \partial_{xx} y = g(y), \qquad g(y) = f(H^{-1}(y))h^2(H^{-1}(y)). \tag{6.10}$$

Since $f$ is defined on $[0,1]$, $g$ is also defined on $[H(0), H(1)] = [0,1]$. Because of (4.9),

$$g(0) = g(H(0)) = 0, \quad g(1) = g(H(1)) = 0, \quad g \text{ has the same sign as } f \circ H^{-1}.$$

Hence if $f$ is monostable then $g$ is monostable. If $f$ is bistable with $f(\theta) = 0$ for some $\theta \in (0,1)$, then $g$ is also bistable with $g(H(\theta)) = 0$, and $H(\theta) \in (H(0), H(1)) = (0,1)$.

We compute

$$g'(y) = f'\big(H^{-1}(y)\big) + 2f(H^{-1}(y))\frac{h'(H^{-1}(y))}{h(H^{-1}(y))}.$$

In particular, $g'(0) = f'(0)$.

Obviously, if there exists a traveling wave for (6.10), $y(t,x) = \widetilde{y}(x - ct)$, connecting 1 to 0, then $p(t,x) := H^{-1}\big(H(0) + (H(1) - H(0))\widetilde{y}(x - ct)\big)$ is a traveling wave for (6.3), connecting 1 to 0.

Then we can compare the wave speeds for (6.10) and for (6.1).

1. If $f$ is monostable, then there exists a minimal traveling speed $c_*$. such that for all $c \geq c_*$, there exists a unique, decreasing, traveling wave $0 \leq y \leq 1$ for (6.10), connecting 1 to 0. Moreover, if KPP condition $g(x) \leq g'(0)x$ holds on $[0,1]$, then $c_* = 2\sqrt{g'(0)} = 2\sqrt{f'(0)}$.

   We notice that the KPP condition $g(x) \leq g'(0)x$ for all $x \in (0,1)$ holds if and only if $f(z)h^2(z) \leq H(z)f'(0)$, by setting $z = H^{-1}(x)$. Hence if $f$ itself satisfies the KPP condition, *i.e.* satisfies $f(z) \leq f'(0)z$, it suffices to check $h^2(z) \leq \frac{H(z)-H(0)}{z}$, $\forall\, z \in (0,1)$. This condition is equivalent to concavity of $H$ on $(0,1)$, *i.e.* $h' \leq 0$ on $(0,1)$.

2. If $f$ is bistable, then there exists a unique traveling wave $(c_*, v)$ for (6.10), decreasing, connecting 1 to 0 and $c_* < 0$ if $G(1) < 0$, $\quad c_* = 0$ if $G(1) = 0$, $\quad c_* > 0$ if $G(1) > 0$, where $G(1) = \int_0^1 g(v)\,dv$ (see [187]). Using the definition of $g$ in (6.10) we get

$$G(1) = \int_0^1 g(y)dy = \int_0^1 f(x)h^4(x)dx.$$

□

**Remark 6.2.** *If $h \equiv 1$ then $H = Id$ and we recover $f = g = \widetilde{g}$.*

**Remark 6.3.** *In the monostable case we find $c_* = 2\sqrt{f'(0)}$, so the minimal speed for (6.3) and for (6.1) are the same.*

If $f$ is bistable and $G(1) > 0$, the sharp threshold property (see [163]) applies to equation (6.10), hence to equation (6.3).

### 6.4.3 Critical propagule size

To identify the initial data that induce invasion, we can compute "propagules" (also called "bubbles"), that is, compactly supported subsolutions to the parabolic problem (6.3). This was stated in Proposition 6.1, that we are going to prove below.

The concept of critical propagule size, that is the minimal "size" of an initial data to ensure invasion, was studied in [29]. We reproduce here for equation (6.10) the computations that can be found in [29] and [212], and deduce an expression of the critical propagule for equation (6.3).

*Proposition 6.1.* We introduce the following Cauchy system associated with (6.3)

$$\begin{cases} p'' + 2\dfrac{h'(p)}{h(p)}(p')^2 + f(p) = 0, & \text{on } [0,+\infty) \\ p(0) = \alpha, \qquad p'(0) = 0 \end{cases} \tag{6.11}$$

Multiplying equation (6.11) by $h(p)^2$ yields $\left(h(p)^2p'\right)' = -f(p)h(p)^2$. Then, multiplying by $h(p)^2p'$ and integrating over $[0,x)$ yields

$$\frac{1}{2}\Big(\big(h(p)^2p'\big)^2 - \big(h(\alpha)^2p'(0)\big)^2\Big) = -\mathcal{F}(p) + \mathcal{F}(p(0)),$$

where $\mathcal{F}$ is an antiderivative of $p \mapsto f(p)h(p)^4$.

We are looking for a decreasing solution $p$ on $[0,+\infty)$. Since $p'(0) = 0$ we get

$$p' = -\frac{\sqrt{2(\mathcal{F}(\alpha) - \mathcal{F}(p))}}{h(p)^2}.$$

Note that since $h(p)^4 > 0$, $\mathcal{F}'$ has the same sign as $f$. If $h$ is constant, we recover the case of equation (6.3) without correction term.

We make a change of variable and check that $v_\alpha := \max(p, 0)$ has support equal to $[0, L_\alpha]$ where

$$L_\alpha := \int_0^\alpha \frac{h(p)^2}{\sqrt{2(\mathcal{F}(\alpha) - \mathcal{F}(p))}}\,dp. \tag{6.12}$$

As for the "classical case" (without $h$) treated in [212], convergence of this integral is straightforward (recalling $\alpha > \theta$). Thus $L_\alpha < \infty$.

Hence we constructed a family $(v_\alpha)_{\theta_c<\alpha<1}$ of compactly supported sub-solutions, where $0 \leq v_\alpha \leq \alpha$. □

## 6.5 Proofs for the heterogeneous case: blocking waves and barrier sets

This section is devoted to the proof of the main results concerning existence of blocking fronts, *i.e.* Theorem 6.2 and Proposition 6.3. This proof is divided in several steps.

The general strategy is to derive as much as possible from the comparison principle alone. This allows for a simple description of the barrier set as positively invariant, and of barriers themselves as decreasing profiles. This is the object of Subsection 6.5.1, where we prove Proposition 6.2 and first point of Proposition 6.3.

Then, we adopt the viewpoint of an equivalent (double-)shooting problem. This is done in Subsection 6.5.2, and enables us to prove that the gradient strength $C$ must be greater than the speed of the traveling wave solution to the homogeneous problem $c_*(f)$.

In order to study this shooting problem, we use on the first hand a phase-plane method, developed in Subsection 6.5.3. This method proves that barriers can be compared to each other, and also provides useful ingredients for the remainder of the proofs.

On the second hand, we get both qualitative (monotonicity) and quantitative (limiting values) results for the solutions of the double-shooting problem without using the phase-plane method, but rather from direct computations, in Subsection 6.5.4.

Then the main results (Theorem 6.2 and Proposition 6.4) are proved in Subsection 6.5.5. This Subsection is the most substantial one, because we gather the ideas and results of the two previous Subsections, introduce a Wronskian argument and also state an additional result on the behavior of barriers as $C$ goes to $+\infty$ (Lemma 6.7).

The final Proposition 6.3 is proved directly from results of Subsection 6.5.6, and the extension to non-constant gradients (Corollary 6.1), which relies simply on a comparison principle, is proved in Subsection 6.5.7.

### 6.5.1 Preliminaries

In this first Subsection we mainly use the comparison principle and sub- and supersolutions method from Section 6.3. We prove here that the barriers (see Definition 6.1) are decreasing. This is the first point of Proposition 6.3.

**Lemma 6.1.** *If $(C, L) \in \mathcal{B}(f)$ and $p$ is a $(C, L)$-barrier, then $p$ is decreasing.*

*Proof.* For any $x \in (-\infty, -L]$, we have

$$\frac{1}{2}p'(x)^2 + F(p(x)) = F(1).$$

Hence $p' = 0$ if and only if $p(x) = 1$, but the maximum principle forbids it (1 is a super-solution so $p$ cannot touch it).

Similarly, $p'$ does not change its sign on $[L, +\infty)$, except possibly if $p = \theta_c$ or $p = 0$. $p = 0$ is impossible by the same argument as before. Assume $p(L) < \theta_c$. Then:

$$\frac{1}{2}p'(L)^2 + F(p(L)) = F(0) = 0.$$

In addition we claim $p'(L) < 0$. To prove this last fact we introduce

$$x_m := \inf\{x > -L, p'(x) = 0\}.$$

By contradiction, we assume $x_m < L$. There are two possibilities.

Either $p(x_m) < \theta_c$. In this case, $\frac{1}{2}p'(x_m)^2 + F(p(x_m)) < 0$. Since $\psi : x \mapsto \frac{1}{2}p'(x)^2 + F(p(x))$ is decreasing and is equal to 0 at $x = L$, this contradicts $x_m < L$. (Indeed, for all $x \in (-L, L)$, $\psi(x) = F(1) - C_N \int_{-L}^{x} p'(x')^2 dx'$.) Or $p(x_m) \geq \theta_c$. If $1 > p(x_m) \geq \theta_c$ then $-p''(x_m) = -p''(x_m) - Cp'(x_m) = f(p(x_m)) > 0$, hence $p$ reaches a local maximum at $x_m$, which is absurd because this contradicts the definition of $x_m$. Hence $p' < 0$ on $[-L, L]$.

Because $0 \leq p \leq 1$ and because of its limits at $\pm\infty$, $p$ is necessarily decreasing on $(-\infty, -L] \cup [L, +\infty)$. □

Existence of a barrier means that the (logarithmic) gradient of total population is enough to stop the bistable propagation. On the contrary, when there is no barrier, then bistable propagation takes place. This is the object of Proposition 6.2, which we prove below.

*Proposition 6.2.* The first point comes directly from the comparison principle (Proposition 4.6), since $p_B$ is a stationary solution, hence a super-solution to (6.4). It is easily checked that $p_B < 1$ by considering a maximum of this function.

First, assume $(C, L) \in \mathcal{B}(f)$ and $p^0 > p_B$ for the maximal barrier $p_B$. By hypothesis, it is unstable from above, hence there exists a sub-solution $\phi$ to (6.6) between $p_B$ and $p^0$. Hence by the comparison principle $p(t, \cdot)$ is bounded from below by $p_\phi(t, \cdot)$, for all $t \geq 0$, where $p_\phi$ is the solution to (6.4) with initial datum $\phi$. Since $p_\phi$ is increasing in $t$ (because initial datum is a subsolution), it converges to some $p^*_\phi$ as $t \to \infty$. However, $p^*_\phi$ is a solution to (6.6) with the last hypotheses on $p(\pm\infty)$ relaxed. Because $p_B$ is a maximal barrier (there is no element above it), $p^*_\phi(-\infty) = 1$ and $p^*_\phi > p_B$, this implies that $p^*_\phi(+\infty)$ is a zero of $f$ which is not 0, hence it must be either $\theta$ or 1. By contradiction, assume $p^*_\phi(+\infty) = \theta$. Then by monotonicity $p^*_\phi \geq \theta$ and thus $-(p^*_\phi)'' = f(p^*_\phi) \geq 0$, so $p^*_\phi$ is concave. Because $p^*_\phi$ is also decreasing, we get the contradicting conclusion that $p^*_\phi$ cannot have a finite limit at $+\infty$. Hence $p^*_\phi(+\infty) = 1$ and thus $p^*_\phi \equiv 1$.

Finally, if $(C, L) \notin \mathcal{B}(f)$, because $p_0 > p_B$ or $\lim_{-\infty} p_0 = 1$, we can always pick a sub-solution $\phi$ which is below $p^0$. For example, a translated $\alpha$-bubble (from Proposition 6.1 in the case $h = 0$) $v_\alpha(\cdot - \tau)$ for some $\tau > 0$ large enough. The solution to (6.4) with initial datum $\phi$, say $p_\phi(t, \cdot)$ is increasing in $t$, and by the comparison principle it is below $p$ for all $t$. Because it is increasing, its limit as $t \to \infty$ is well-defined and it is a solution to (6.6) without the final conditions (on $p(\pm\infty)$). Since (6.6) has no solution, this implies that $p_\phi(t, \cdot) \to 1$. Hence $p \to 1$. □

To simplify notably the study of the barrier set $\mathcal{B}(f)$, we obtain a simple positivity property.

**Proposition 6.5.** *For all $B_1 \in \mathcal{B}(f)$ and $B_2 \in [0, +\infty)^2$, $B_1 + B_2 \in \mathcal{B}(f)$.*

*Proof.* Let $B_1 = (C_1, L_1)$, $p_1$ be a solution to (6.6) where $C = C_1$ and $L = L_1$. Let $B_2 = (C_2, L_2)$. Then, $p_1$ is *decreasing* (by Lemma 6.1), hence

$$\begin{aligned} -p_1'' - (C_1 + C_2)p_1' \geq p_1'' - C_1 p_1' = f(p_1) \text{ on } [-L_1, L_1], \\ -p_1'' - (C_1 + C_2)p_1' \geq p_1'' = f(p_1) \text{ on } [-(L_1 + L_2), -L_1] \bigcup [L_1, L_1 + L_2], \\ -p_1'' = f(p_1) \text{ on } \mathbb{R} \backslash [-(L_1 + L_2), L_1 + L_2]. \end{aligned}$$

In other words, $p_1$ is a supersolution of (6.6) for $C = C_1 + C_2$, $L = L_1 + L_2$.

On the other hand, the $\alpha$-bubbles from Proposition 6.1 give us subsolutions, and we can select any of them. Upon moving it far enough towards $-\infty$, it will be below $p_1$. We simply need to consider $v_\alpha(\cdot - \tau)$ for $\tau > 0$ large enough, which will be the required subsolution.

This implies that we can construct a solution $p$ to (6.6) for $C = C_1 + C_2$ and $L = L_1 + L_2$, lying between the $\alpha$-bubble and $p_1$, by Proposition 4.5. As $p_1$ is decreasing, one could check that $p$ is decreasing as well, and thus it admits limits at $\pm\infty$. Then one could check that $p(+\infty) = 0$ and $p(-\infty) = 1$, whence $p$ is a barrier. Hence $B_1 + B_2 \in \mathcal{B}(f)$. □

### 6.5.2 A double shooting-argument.

To get a better description of $\mathcal{B}(f)$ than allowed by comparison principle alone, we introduce a double shooting-argument. We separate the study of equation (6.6) on $[-L, L]$ by introducing

$$\beta = p(-L), \quad \alpha = p(L).$$

We are left with a slightly differently rephrased problem: given $0 < \alpha < \beta < 1$, we are looking for $C, L > 0$ such that

$$\begin{cases} -p'' - Cp' = f(p), \\ p(-L) = \beta, \quad p(L) = \alpha, \\ \frac{1}{2}p'(-L)^2 + F(\beta) = F(1), \quad \frac{1}{2}p'(L)^2 + F(\alpha) = 0. \end{cases} \tag{6.13}$$

The two equations (6.6) and (6.13) are obviously directly related.

**Proposition 6.6.** *Let $C, L > 0$. If $(C, L) \in \mathcal{B}(f)$, then there exists $(\alpha, \beta)$ such that (6.13) has a solution. Conversely, if there are $\alpha, \beta$ and $C, L$ such that (6.13) has a solution, then its solutions are also solutions to (6.6).*

The proof is a straightforward computation. A first property of (6.13) can easily be proven:

**Proposition 6.7.** *For any $0 < \alpha < \beta < 1$ with $\alpha < \theta_c$, there exists a unique $C = \gamma(\alpha, \beta)$ such that the system (6.13) has a solution, associated with a unique $L = \lambda(\alpha, \beta)$.*

*Proof.* Here we employ a shooting argument. Let $p_\alpha$ be the unique (by Cauchy-Lipschitz theorem), decreasing (by similar arguments as in Lemma 6.1) solution to

$$\begin{cases} -p''_\alpha - Cp'_\alpha = f(p_\alpha), \\ p_\alpha(L) = \alpha, \quad \frac{1}{2}p'_\alpha(L)^2 + F(\alpha) = 0. \end{cases} \tag{6.14}$$

Because $p_\alpha$ is decreasing, we can introduce $X_\alpha : [p_\alpha(L), p_\alpha(-L)] \to [-L, L]$ such that $p_\alpha(X_\alpha(p)) = p$. Using the method of [30] we also introduce $w_\alpha(p) := \frac{1}{2}p'_\alpha(X_\alpha(p))^2 + F(p)$. Then:

$$\begin{cases} w'_\alpha(p) = C\sqrt{2\big(w_\alpha(p) - F(p)\big)}, \\ w_\alpha(\alpha) = 0. \end{cases} \tag{6.15}$$

The solution of this problem exists as long as $w_\alpha(p) \geq F(p)$. For $\alpha < \theta_c$, since $F(p) < 0$ for $p \in (0, \theta_c)$, we deduce that the solution exists at least on $(\alpha, \theta_c)$. Let us denote $p_0 \leq 1$ such that $(\alpha, p_0)$ is the maximum interval in $(\alpha, 1)$ of existence of a solution to (6.15). We have $p_0 \geq \theta_c$.

Then, let $\beta > \alpha$. We are going to show that we can choose $C$ such that $w_\alpha(\beta) = F(1)$. We first notice that on $(\alpha, \theta_c)$, we have $F(p) < 0$ thus $w'_\alpha(p) > C\sqrt{2w_\alpha(p)}$. It implies that $w_\alpha(p) > \frac{1}{2}C^2(p-\alpha)^2$ on $(\alpha, \theta_c)$. Thus if $C$ is large enough, surely we will have $w_\alpha(\beta) > F(1)$.

Conversely, we have $w'_\alpha(p) \leq C\sqrt{2(w_\alpha(p) - F(\theta))}$, since $F(\theta) = \min_{[0,1]} F$. Integrating on $(\alpha, p)$, we deduce $w_\alpha(p) \leq F(\theta) + \big(\frac{1}{\sqrt{2}}C(p-\alpha) + \sqrt{-F(\theta)}\big)^2$. Thus we may choose $C$ small enough such that $w_\alpha(\beta) < F(1)$. Finally, by differentiating (6.15) with respect to $C$, we deduce that the solution $w$ is increasing with respect to $C$.

Hence for each $\beta$ there exists a unique $C = \gamma(\alpha, \beta)$ such that $w_\alpha(\beta) = F(1)$. We rename this solution as $w_{\alpha,\beta}$, so that

$$\begin{cases} w'_{\alpha,\beta}(p) = \gamma(\alpha, \beta)\sqrt{2\big(w_{\alpha,\beta}(p) - F(p)\big)}, \\ w_{\alpha,\beta}(\alpha) = 0,\ w_{\alpha,\beta}(\beta) = F(1). \end{cases} \tag{6.16}$$

To retrieve the value of $L$, such that $w_{\alpha,\beta}$ comes from a $p_\alpha$ solution of (6.14) with $p_\alpha(-L) = \beta$, $\frac{1}{2}\big(p'_\alpha(-L)\big)^2 + F(p_\alpha(-L)) = F(1)$, we simply have to remark that $L = \frac{1}{2}\int_\beta^\alpha \big(X_\alpha\big)'(p)dp$. To compute it from $w_{\alpha,\beta}$ we notice that $(X_\alpha)'(p) = 1/p'_\alpha\big(X_\alpha(p)\big)$. Hence we define

$$\lambda(\alpha, \beta) := \frac{1}{2}\int_\alpha^\beta \frac{1}{\sqrt{2\big(w_{\alpha,\beta}(p) - F(p)\big)}}dp. \tag{6.17}$$

(Indeed, recall that $p' < 0$ on $(-L, L)$). Then $L = \lambda(\alpha, \beta)$ is uniquely defined.

□

**Lemma 6.2.** *Functions $\gamma$ and $\lambda$ defined in Proposition 6.7 are continuous on $\{(\alpha, \beta), 0 < \alpha < \theta_c,$ and $\alpha < \beta < 1\}$.*

*Proof.* We transform problem (6.16) into a ordinary differential equation $w'(p) = \gamma J(w(p), p)$, with either $w(\alpha) = 0$ or $w(\beta) = F(1)$, and $\gamma > 0$.

On the prescribed set for $\alpha, \beta$, the function $J$ is uniformly Lipschitz along any forward trajectory. This implies the continuity of $w$ with respect to $\gamma$, and finally the continuity of $\gamma$ with respect to $\beta$ (in the case when we impose $w(\alpha) = 0$), and with respect to $\alpha$ (when we impose $w(\beta) = F(1)$).

This implies the continuity of $\lambda$. □

At this stage we can already get a simple consequence of the shooting viewpoint:

**Proposition 6.8.** *Let $L > 0$. If $(C, L) \in \mathcal{B}(f)$ then $C > c_*(f)$.*

*Proof.* This comes from the fact that there exists $w_{0,1}$ such that

$$\begin{cases} w'_{0,1} = c_*(f)\sqrt{2(w_{0,1} - F)}, \\ w_{0,1}(0) = 0,\ w_{0,1}(1) = F(1). \end{cases} \tag{6.18}$$

And the associated $\lambda(0,1)$ is equal to $+\infty$. By comparison of solutions to (6.16), no $(\alpha, \beta) \neq (0,1)$ could give a $w_{\alpha,\beta}$ associated with $C \leq c_*(f)$. □

### 6.5.3 A graphical digression on phase plane analysis.

Equation (6.13) can be easily interpreted in the phase plane $(p, p')$. In addition, phase plane arguments allow us to study the structure of the barrier set in detail, and are necessary to solve the double-shooting problem in Subsection 6.5.5 below. For this interpretation, we follow the presentation of [148]. Let $X = p$, $Y = p'$. Equation (6.13) rewrites into the system

$$\begin{cases} X' = Y, \quad X(0) = X_0, \\ Y' = -CY - f(X), \quad Y(0) = Y_0. \end{cases} \tag{6.19}$$

The energy $E : \mathbb{R}^2 \to \mathbb{R}$ may be defined as

$$E(X,Y) := \frac{1}{2}Y^2 + F(X). \tag{6.20}$$

Two interesting curves appear:

$$E^{-1}\big(F(1)\big) \supset \Gamma_B := \Big\{(x,y) \in [0,1] \times (-\infty, 0],\ y = -\sqrt{2\big(F(1) - F(x)\big)}\Big\}, \tag{6.21}$$

$$E^{-1}\big(0\big) \supset \Gamma_A := \Big\{(x,y) \in [0,\theta_c] \times (-\infty, 0],\ y = -\sqrt{-2F(x)}\Big\}. \tag{6.22}$$

Indeed, a $(C,L)$-barrier can be seen there as a trajectory of (6.19) such that $\big(X(L), Y(L)\big) \in \Gamma_A$, with $\big(X(-L), Y(-L)\big) \in \Gamma_B$.

Therefore, we are left studying the image of $\Gamma_B$ by the flow of (6.19), which we denote by $\phi_t^C : \mathbb{R}^2 \times \mathbb{R}^2$, at time $t$.

**Lemma 6.3.** *The energy decreases along trajectories:*

$$\frac{d}{dt}E\big(X(t), Y(t)\big) = -CY(t)^2.$$

*At the three equilibrium points of the system it is equal to:*

$$E(0,0) = 0,\ E(\theta, 0) = F(\theta) < 0,\ E(1,0) = F(1) > 0.$$

*It is therefore minimal at* $(\theta, 0)$.

This is a straightforward computation.

Let $\chi \in [\theta_c, 1]$. We define the level set of $E$

$$\Gamma_\chi := E^{-1}\big(F(\chi)\big) = \Big\{(x,y) \in [0,\chi] \times (-\infty, 0], y = -\sqrt{2\big(F(\chi) - F(x)\big)}\Big\}.$$

Note that $\Gamma_1 = \Gamma_A$ and $\Gamma_{\theta_c} = \Gamma_B$, by definition.

For $\chi \in [\theta_c, 1]$ and $P \in \Gamma_\chi$, let $\nu_\chi(P)$ be the inward normal vector ("inward" meaning pointing towards $y = 0$). Then we claim

**Lemma 6.4.** *For all* $\chi \in [\theta_c, 1]$, $P \in \Gamma_\chi$, $C > 0$, *the flow of* (6.19) *is inward:* $\frac{d}{dt}\phi_0^C(P) \cdot \nu_A(P) > 0$.

*Proof.* First, system (6.19) may be rewritten $\dot{u} = G(u), u(0) = u_0$, where $u = (X,Y)$ and $u_0 = (X_0, Y_0)$. Then, $\frac{d}{dt}\phi_0^C(u_0) = G(u_0)$, obviously (and similarly, $\frac{d}{dt}\phi_t^C(u_0) = G(u(t))$). Now, recall that $\Gamma_\chi = \{(\alpha, -\sqrt{2\big(F(\chi) - F(\alpha)\big)}$ where $0 \leq \alpha \leq \chi\}$.

Hence if $P = (\alpha, -\sqrt{2(F(\chi) - F(\alpha))})$,

$$\nu_\chi(P) = \begin{pmatrix} -\dfrac{f(\alpha)}{\sqrt{2(F(\chi) - F(\alpha))}} \\ 1 \end{pmatrix}$$

and

$$\frac{d}{dt}\phi_0^C(P) = G(p) = \begin{pmatrix} -\sqrt{2(F(\chi) - F(\alpha))} \\ C\sqrt{2(F(\chi) - F(\alpha))} - f(\alpha) \end{pmatrix}.$$

Hence

$$D\phi_0^C(P) \cdot \nu_\chi(P) = C\sqrt{2(F(\chi) - F(\alpha))} > 0.$$

□

The following crucial property will make us able to show that barriers are ordered. Its graphical interpretation is shown on Figure 6.7.

**Lemma 6.5.** *Let $p_1, p_2 \in (0,1)$ with $p_1 < p_2$. We denote $(X_1, Y_1)$ (resp. $(X_2, Y_2)$) the unique solution of (6.19) with $X_1(0) = p_1$ (resp. $X_2(0) = p_2$) and $Y_1(0) = -\sqrt{2\big(F(1) - F(p_1)\big)}$ (resp. $Y_2(0) = -\sqrt{2\big(F(1) - F(p_2)\big)}$.*

*Let $t_M > 0$ be such that for all $t < t_M$, $Y_1, Y_2 < 0$, $X_1, X_2 > 0$. Then*

$$\forall t < t_M, \quad X_1(t) < X_2(t). \tag{6.23}$$

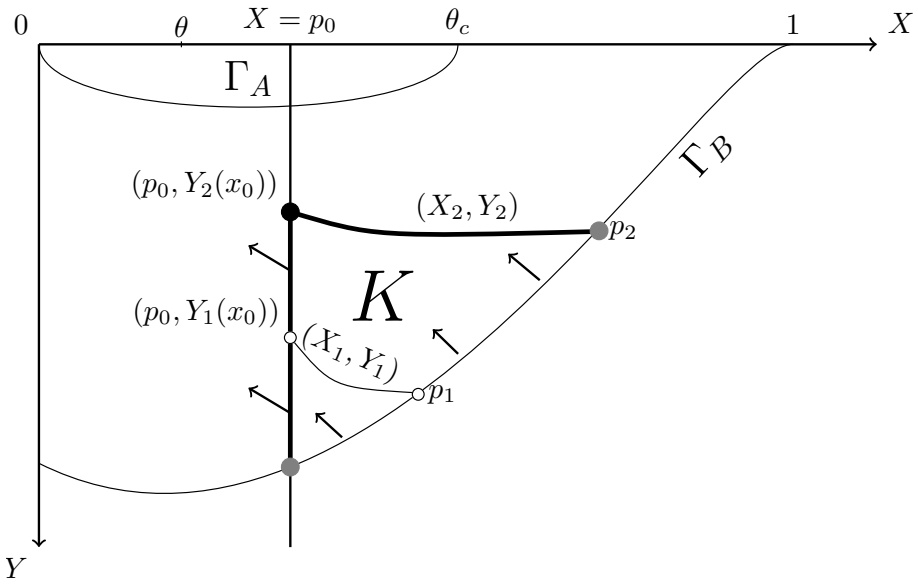



Figure 6.7: Sketch of the phase-plane argument in the proof of Lemma 6.5. Because the trajectories satisfy $\dot{X} = Y$, this picture is impossible. On the other hand, $Y_1(x_0) > Y_2(x_0)$ would imply that the two trajectories cross each other, which is impossible as well. Whence the claim.

*Proof.* To prove this we introduce

$$t_0 := \inf\{t > 0,\ X_1(t) = X_2(t)\}.$$

If $t_0 = +\infty$, we are done. If $t_0 < +\infty$, we first note that if $t < t_0$ then $X_1(t) < X_2(t)$, by definition of $t_0$ and continuity of $X_1, X_2$. As a consequence, $\frac{d}{dt}(X_2 - X_1)(t_0) \leq 0$, and $Y_1(t_0) \geq Y_2(t_0)$.

We show that phase-plane reasoning imposes

$$Y_1(t_0) \leq Y_2(t_0).$$

To prove this fact, we first observe that (6.19) has its flow from the right to the left along any vertical line ($X$ = constant), in the quadrant $X > 0, Y < 0$ (because $\dot{X} = Y$).

Moreover, $Y_2(t_0) > -\sqrt{2\big(F(1) - F(t_0)\big)}$, because $E(X_2(t_0), Y_2(t_0)) < F(1) = E(X_2(0), Y_2(0))$, by Lemma 6.3 ($E$ was defined in (6.20)).

Hence the trajectory of $(X_1, Y_1)$ enters at $x = 0^+$ the compact set $K$ defined by the vertical line $X = X_1(t_0)$, the trajectory of $(X_2, Y_2)$ and $\Gamma_B$ (that is, the level set $F(1)$ of $E$). Indeed, $(X_1(0), Y_1(0))$ is on the part of $\Gamma_B$ which defines the border of $K$, and the flow of (6.19) is inward at this point (by Lemma 6.4).

Moreover the trajectory of $(X_1, Y_1)$ cannot exit $K$ but on the line $X = X_1(x_0) =: p_0$: its energy decreases and it cannot cross the trajectory of $(X_2, Y_2)$. More precisely, it exits $K$ on the segment

$$\big[\big(p_0, -\sqrt{2(F(1) - F(t_0))}\big), \quad \big(p_0, Y_2(t_0)\big)\big] \subset \{X = p_0\}.$$

As a consequence, $Y_1(t_0) \le Y_2(t_0)$.

Hence $Y_1(t_0) = Y_2(t_0)$, which contradicts the uniqueness of the solutions of (6.19) (since $X_1(t_0) = X_2(t_0)$). Finally, $t_0 = +\infty$ and Lemma 6.5 is proved. □

### 6.5.4 Back to the double-shooting.

Thanks to the double-shooting argument (6.5.2), determining $\mathcal{B}(f)$ amounts to computing the image of $\{0 < \alpha < \beta < 1,\ \alpha < \theta_c\}$ by $(\gamma, \lambda)$ defined in Proposition 6.7. Even without the phase-plane viewpoint from the previous Subsection, we can prove that these functions $\gamma, \lambda$ have nice monotonicity properties, compute their limits, and already get a good description of the barrier set.

**Proposition 6.9.** *Let $\gamma$ and $\lambda$ be defined as in Proposition 6.7 on the set $\{(\alpha, \beta) \in (0,1)^2, \quad 0 \le \alpha \le \theta_c,\ \beta > \alpha\}$. $\gamma(\alpha, \beta)$ is increasing in $\alpha$, decreasing in $\beta$. $\lambda(\alpha, \beta)$ is increasing in $\beta$.*

*Proof.* Take $0 < \alpha < \beta$ with $\alpha < \theta_c$, $C = \gamma(\alpha, \beta)$ and $w$ be the solution of (6.16) associated with $C$ and $\beta$. Similarly, take $\tilde{\beta} > \beta$ and let $\tilde{C} := \gamma(\alpha, \tilde{\beta})$ and $\tilde{w}$ the solution of (6.16) associated with $\tilde{C}$ and $\tilde{\beta}$ (*i.e.* $\tilde{w}(\tilde{\beta}) = F(1)$). Assume by contradiction that $\tilde{C} \ge C$. Then $\tilde{w}$ is a supersolution of the equation satisfied by $w$, with initial datum $\tilde{w}(\alpha) = 0$. Hence $\tilde{w} \ge w$ on $[\alpha, \beta]$ and $\tilde{w}(\beta) \ge F(1) = \tilde{w}(\tilde{\beta})$. This is a contradiction since $\tilde{w}$ is increasing.

Hence, $\tilde{C} < C$ and thus, as $w(\alpha) = \tilde{w}(\alpha) = 0$, one gets $\tilde{w} < w$ on $(\alpha, \beta)$. We can therefore compute

$$2\lambda(\alpha, \tilde{\beta}) = \int_\alpha^{\tilde{\beta}} \frac{dx}{\sqrt{2\big(\tilde{w}(x) - F(x)\big)}} > \int_\alpha^{\beta} \frac{dx}{\sqrt{2\big(w(x) - F(x)\big)}} = 2\lambda(\alpha, \beta),$$

proving the monotonicity of $\lambda$ as a function of $\beta$.

The monotonicity of $\gamma$ with respect to $\alpha$ is proved similarly. □

In addition, some limits of $\gamma$ and $\lambda$ can be computed directly.

**Proposition 6.10.** *Functions $\gamma, \lambda$ satisfy: $\gamma(\alpha, \beta) \to +\infty$ as $\beta \searrow \alpha$.*
*$\lambda(\alpha, \beta) \to +\infty$ as $\beta \to 1$, $\lambda(\alpha, \beta) \to +\infty$ as $\alpha \to 0$. $\lambda(\alpha, \beta) \to 0$ as $\beta - \alpha \to 0$.*

*Proof.* We have already proved in Proposition 6.7 that

$$w(p) \le F(\theta) + \big(\frac{1}{\sqrt{2}}\gamma(\alpha, \beta)(p - \alpha) + \sqrt{-F(\theta)}\big)^2.$$

Hence, taking $p = \beta$, one has $F(1) - F(\theta) \le \big(\frac{1}{\sqrt{2}}\gamma(\alpha, \beta)(\beta - \alpha) + \sqrt{-F(\theta)}\big)^2$. If $\gamma(\alpha, \beta)$ does not diverge to $+\infty$ when $\beta \searrow \alpha$, this function would be bounded since it is monotonic, and thus, passing to the limit in the inequality: $F(1) - F(\theta) \le \big(\sqrt{-F(\theta)}\big)^2 = -F(\theta)$, this would contradict $F(1) > 0$.

Now, the function $\gamma(\alpha, \cdot\,)$ being decreasing and bounded from below by $c^*$, it converges to some limit $C^\infty$ as $\beta \nearrow 1$. As $\lambda(\alpha, \cdot\,)$ is increasing, if it does not diverge to $+\infty$ then it converges to some limit $\lambda^\infty$. We could thus derive a solution $p$ of

$$\begin{cases} -p'' - C^\infty p' = f(p), \text{ on } (-\lambda^\infty, 0), \\ \frac{1}{2}(p'(-\lambda^\infty))^2 + F(1) = F(1), \\ \frac{1}{2}(p'(\lambda^\infty))^2 + F(\alpha) = 0. \end{cases}$$

This implies $p'(-\lambda^\infty) = 0$ and thus $p \equiv 1$ by uniqueness, which contradicts $\frac{1}{2}(p'(0))^2 + F(\alpha) = 0$.

The convergence of $\lambda(\,\cdot\,, \beta)$ when $\alpha \to 0$ is proved similarly.

Finally, we know that $w_{\alpha,\beta}(p) - F(p) \geq -\min_{[\alpha,\beta]} F$, since $w_{\alpha,\beta} \geq 0$. Hence if $\beta$ is close enough to $\alpha$, $w_{\alpha,\beta}(p) - F(p) \geq -\frac{1}{2}F(\alpha)$ (uniformly in $\beta$). Then,

$$2\lambda(\alpha,\beta) = \int_\alpha^\beta \frac{dp}{\sqrt{2w_{\alpha,\beta}(p) - F(p)}} \leq \frac{\beta - \alpha}{\sqrt{-F(\alpha)}}$$

As $\beta \to \alpha$, we deduce that $\lambda(\alpha,\beta) \to 0$, and similarly when $\alpha \to \beta \in (0, \theta_c)$. □

**Lemma 6.6.** *For all $\alpha \in (0,\theta_c)$, $\beta \in (\alpha, 1)$,*

$$2\lambda(\alpha,\beta)\gamma(\alpha,\beta) \geq 1 - \sqrt{\frac{-F(\theta)}{F(1) - F(\theta)}}. \tag{6.24}$$

*Moreover, for $0 < \beta < \theta_c$, we have*

$$\lim_{\alpha \to \beta_-} 2\lambda(\alpha,\beta)\gamma(\alpha,\beta) = \frac{1}{2}\ln\left(1 - \frac{F(1)}{F(\beta)}\right). \tag{6.25}$$

*Proof.* The estimate from below is only based on the following inequalities

$$F(1) \geq w_{\alpha,\beta}(p) \geq F(p) \geq F(\theta).$$

They imply, as stated before (in the proof of Proposition 6.10):

$$\gamma(\alpha,\beta) \geq \frac{\sqrt{2}}{\beta - \alpha}\left(\sqrt{F(1) - F(\theta)} - \sqrt{-F(\theta)}\right).$$

Moreover, $\sqrt{w_{\alpha,\beta}(p) - F(p)} \leq \sqrt{F(1) - F(\theta)}$. Thus,

$$2\lambda(\alpha,\beta) \geq (\beta - \alpha)\frac{1}{\sqrt{2\big(F(1) - F(\theta)\big)}}. \tag{6.26}$$

Combining these estimates yields (6.24).

Let us fix $\beta \in (0,\theta_c)$, for $0 < \alpha < \beta$, we have, using (6.17) and (6.16),

$$2\lambda(\alpha,\beta)\gamma(\alpha,\beta) = \int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(x))}\,dx.$$

On the one hand, we have

$$\int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(x))}\,dx - \int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(\beta))}\,dx = \int_\alpha^\beta \frac{w'(x)}{2}\frac{F(x) - F(\beta)}{(w(x) - F(x))(w(x) - F(\beta))}\,dx.$$

For any $0 < \alpha < \beta < \theta_c$, we have $0 \leq w(x) \leq F(1)$ then

$$\frac{|F(x) - F(\beta)|}{(w(x) - F(x))(w(x) - F(\beta))} \leq \frac{|F(x) - F(\beta)|}{F(x)F(\beta)} \leq \left|\frac{1}{F(\beta)} - \frac{1}{F(x)}\right|.$$

Then, for $\alpha$ close enough to $\beta$, we have

$$\begin{aligned}\left|\int_\alpha^\beta \frac{w'(x)}{2}\frac{F(x) - F(\beta)}{(w(x) - F(x))(w(x) - F(\beta))}\,dx\right| &\leq \int_\alpha^\beta \frac{w'(x)}{2}\,dx\left|\frac{1}{F(\beta)} - \frac{1}{F(\alpha)}\right| \\ &= \frac{F(1)}{2}\left|\frac{1}{F(\beta)} - \frac{1}{F(\alpha)}\right|.\end{aligned}$$

We deduce that

$$\int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(x))}\,dx - \int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(\beta))}\,dx \to 0, \qquad \text{as } \alpha \to \beta_-.$$

On the other hand, we compute

$$\int_\alpha^\beta \frac{w'(x)}{2(w(x) - F(\beta))}\,dx = \frac{1}{2}\ln\left(1 - \frac{F(1)}{F(\beta)}\right).$$

Combining these last identities allows to recover (6.25). □

**Proposition 6.11.** *For all $\epsilon > 0$ small enough, there exists $\alpha_\epsilon < \beta_\epsilon$ with*

$$\gamma(\alpha_\epsilon, \beta_\epsilon) = c_*(f) + \epsilon,$$

*and $\alpha_\epsilon \to 0$, $\beta_\epsilon \to 1$ as $\epsilon \to 0$. Moreover, $\lambda(\alpha_\epsilon, \beta_\epsilon) \xrightarrow{\epsilon \to 0} +\infty$.*

*Proof.* The limit of $\gamma(\alpha, \beta)$ as $\alpha \to 0$ and $\beta \to 1$ exists because of the monotonicity properties of Proposition 6.9. Moreover, $\gamma(\alpha, \beta)$ is bounded from below by $c_*(f)$. Simultaneously, we know that $\lambda(\alpha, \beta) \to +\infty$ as $\alpha \to 0$ and $\beta \to 1$ by Proposition 6.10.

The uniqueness of the bistable traveling wave and continuity of $\gamma$ (Lemma 6.2) imply that

$$\lim_{\alpha \to 0, \beta \to 1} \gamma(\alpha, \beta) = c_*(f).$$

Indeed, let $c$ be this limit. At the limit ($w_{\alpha,\beta}$ and its derivative being uniformly bounded), we get a solution of

$$\begin{cases} w' = c\sqrt{2(w - F)} \\ w(0) = 0,\ w(1) = F(1). \end{cases}$$

This exists if and only if $c = c_*(f)$, by uniqueness of the traveling wave solution to the bistable reaction-diffusion equation. These facts imply the existence of $\alpha_\epsilon, \beta_\epsilon$. □

The following fact may be proved using phase-plane (and more precisely Lemma 6.5), but it also enjoys a simple proof using the properties of $\gamma$, which we propose below.

**Proposition 6.12.** *If $\gamma(\alpha_1, \beta_1) = \gamma(\alpha_2, \beta_2)$, then $\alpha_1 < \alpha_2$ if and only if $\beta_1 < \beta_2$.*

*Proof.* Let $C = \gamma(\alpha_1, \beta_1) = \gamma(\alpha_2, \beta_2)$. Assume $\alpha_1 < \alpha_2$. We can compare $w_1 := w_{\alpha_1,\beta_1}$ and $w_2 := w_{\alpha_2,\beta_2}$ because $w_2(\alpha_2) = 0 < w_1(\alpha_1)$ and as long as $w_2 < w_1$ we also get $w_2' < w_1'$. Hence $w_1(\beta_1) - w_2(\beta_1) > w_1(\alpha_1)$. Since $w_1(\beta_1) = F(1)$ we get

$$w_2(\beta_1) < F(1) - w_1(\alpha_1) < F(1).$$

Since $w_2$ is increasing and $w_2(\beta_2) = F(1)$, this implies $\beta_2 > \beta_1$.

Conversely, if $\beta_1 < \beta_2$, we get $w_2(\beta_1) < w_1(\beta_1) = F(1)$ and $w_2' < w_1'$ as long as $w_2 < w_1$. By contradiction assume $\alpha_1 > \alpha_2$. Then we find $w_1(\alpha_1) = 0 < w_2(\alpha_1)$ and by the previous remark $w_2 < w_1$. This is absurd, whence the result. □

### 6.5.5 Advanced properties of the barrier set.

At this stage, by connecting the phase-plane method and the shooting problem we are ready to prove the following description of $\mathcal{B}(f)$, which encompasses Theorem 6.2 and first point of Proposition 6.4.

**Proposition 6.13.** *For all $L > 0$, there exists $C_*(L) > c_*(f)$ such that $(C, L) \in \mathcal{B}(f) \iff C \geq C_*(L)$. For all $C > c_*(f)$, there exists $L_*(C) > 0$ such that $(C, L) \in \mathcal{B}(f) \iff L \geq L_*(C)$.*

*Furthermore, $C_*(L_*(C)) = C$ and $L_*(C_*(L)) = L$.*

*Proof.* By Propositions 6.9 and 6.10, for any $\alpha \in (0, \theta_c)$ and $L > 0$, there exists a unique $\beta_L(\alpha) > \alpha$ such that $\lambda(\alpha, \beta_L(\alpha)) = L$. In particular, $\big(\gamma(\alpha, \beta_L(\alpha)), L\big) \in \mathcal{B}(f)$.

Hence $C_*(L) := \inf\{C > 0,\ (C, L) \in \mathcal{B}(f)\}$ is well-defined and because of Proposition 6.5, if $C > C_*(L)$ then $(C, L) \in \mathcal{B}(f)$. Moreover, $C_*(L) > c_*(f)$ by Proposition 6.8

Let $C > c_*(f)$. Then we claim there exists $\alpha, \beta$ such that $\gamma(\alpha, \beta) = C$. First, for $\epsilon > 0$ small enough, there exists $\alpha_\epsilon$ (close to 0) and $\beta_\epsilon$ (close to 1) such that $\gamma(\alpha_\epsilon, \beta_\epsilon) = c_*(f) + \epsilon$, by Proposition 6.11.

Hence we can find $\alpha_0, \beta_0$ such that $\gamma(\alpha_0, \beta_0) < C$.

Then since $\gamma(\alpha_0, \beta) \to +\infty$ as $\beta \searrow \alpha_0$ (Proposition 6.10) and $\gamma(\alpha_0, \beta)$ is decreasing in $\beta$ (Proposition 6.9), there exists a unique $\beta_C(\alpha_0)$ such that $\gamma(\alpha_0, \beta_C(\alpha_0)) = C$. Like before, $L_*(C) := \inf\{L > 0,\ (C, L) \in \mathcal{B}(f)\}$ fulfills all properties.

Let $\epsilon > 0$. By definition there exists $\alpha_\epsilon, \beta_\epsilon$ such that

$$\gamma(\alpha_\epsilon, \beta_\epsilon) = C, \quad \lambda(\alpha_\epsilon, \beta_\epsilon) = L_*(C) + \epsilon.$$

Up to extraction we pass to the limit $\epsilon \to 0$ (the couple $(\alpha_\epsilon, \beta_\epsilon)$ is in a compact set). Since $\gamma$ and $\lambda$ are continuous, we get $(C, L_*(C)) \in \mathcal{B}(f)$, and $(C_*(L), L) \in \mathcal{B}(f)$ by a similar argument.

Last point boils down to strict monotonicity of $L_*$. The solution $(X(t), Y(t))$ of

$$\begin{cases} \dot{X} = Y, \quad X(0) = \beta, \\ \dot{Y} = -CY - f(X), \quad Y(0) = -\sqrt{2\big(F(1) - F(\beta)\big)} \end{cases}$$

depends smoothly on $C$ and $\beta$, so we write it $\big(X(t; C, \beta), Y(t; C, \beta)\big)$. We note that by definition

$$2L_*(C) = \inf_{\beta \in (0,1)} \inf_{t>0} \big\{t, \quad E\big(X(t; C, \beta), Y(t; C, \beta)\big) = 0\big\}$$

We denote by $(X_C, Y_C)$ (resp. $(X_\beta, Y_\beta)$) its derivative with respect to $C$ (resp. $\beta$).

From now on we only consider solutions such that $Y < 0$, $X \in [0, 1]$, truncating in time if necessary.

Using indifferently the notations $E = E\big(X(t; C, \beta), Y(t; C, \beta)\big) = E(t; C, \beta)$ we find

$$\partial_C E(t) = \frac{\partial E}{\partial C}(t; C, \beta) = Y_C(t)Y(t) + X_C(t)f(X(t)), \tag{6.27}$$

$$\partial_\beta E(t) = \frac{\partial E}{\partial \beta}(t; C, \beta) = Y_\beta(t)Y(t) + X_\beta(t)f(X(t)). \tag{6.28}$$

Let $t_* = L_*(C) = \inf_{\beta \in (0,1)} \inf\{t > 0,\, E(t; C, \beta) = 0\}$, and assume $\beta_*(C) \in (0, 1)$ realizes this infimum. We claim that if $\partial_C E(t_*(C); C, \beta_*(C)) < 0$, then $L_*$ is strictly monotone at $C$.

Indeed, let $t_*, \beta_*$ be minimal such that $E\big(X(t_*), Y(t_*)\big) = 0$ and assume $\partial_C E(t_*) < 0$. For $\epsilon > 0$ small enough, $E(t_*; C + \epsilon, \beta_*) < 0$ by $\partial_C E < 0$. Hence there exists $t'_* < t_*$ such that $E(t'_*; C + \epsilon, \beta_*) = 0$. This yields $L_*(C + \epsilon) \le t'_* < t_* = L_*(C)$, that is strict monotonicity.

To prove $\partial_C E < 0$, we notice that $(X_C, Y_C)$ and $(X_\beta, Y_\beta)$ are solutions to affine differential systems, with the same homogeneous parts.

$$\begin{cases} \dot{X}_C = Y_C, \quad X_C(0) = 0, \\ \dot{Y}_C = -CY_C - Y - X_C f'(X), \quad Y_C(0) = 0, \end{cases} \tag{6.29}$$

and

$$\begin{cases} \dot{X}_\beta = Y_\beta, \quad X_\beta(0) = 1, \\ \dot{Y}_\beta = -CY_\beta - X_\beta f'(X), \quad Y_\beta(0) = \dfrac{f(\beta)}{\sqrt{2\big(F(1) - F(\beta)\big)}}. \end{cases} \tag{6.30}$$

Moreover we notice that $X_\beta(t) > 0$ for all $t \ge 0$. Indeed, because of Lemma 6.5, $X$ is monotone with respect to its boundary data, that is $X_\beta \ge 0$. Then, it suffices to show that $X_\beta$ cannot reach 0 in finite time. This is a straightforward application of Cauchy-Lipschitz theorem: indeed, since $X_\beta \ge 0$, if $X_\beta(t_0) = 0$ for some $t_0 > 0$ then $\dot{X}_\beta(t_0) = 0$, hence $Y_\beta(t_0) = 0$ and finally $(X_\beta, Y_\beta) \equiv (0, 0)$ by Cauchy-Lipschitz theorem.

Then, we compute the differential equation satisfied by the Wronskian $w(t) := Y_C X_\beta - Y_\beta X_C$:

$$\begin{aligned} w'(t) &= \dot{Y}_C X_\beta - \dot{Y}_\beta X_C \\ &= -Cw - Y X_\beta. \end{aligned}$$

Because $Y < 0$ and $X_\beta > 0$ we get

$$\begin{cases} (w' + Cw)(t) \ge 0 \, \forall t, \quad (w' + Cw)(t = 0) > 0, \\ w(0) = 0. \end{cases}$$

Hence if $t > 0$ then $w(t) > 0$. We can then compute $w$ at $(t_*, \beta_*)$. At this point, necessarily

$\partial_\beta E = 0$ (necessary condition for minimality on $\beta$). And $w(t_*) > 0$ is equivalent to

$$\begin{aligned} & Y_C X_\beta > X_C Y_\beta \\ \iff & Y_C > \frac{X_C Y_\beta}{X_\beta} \\ \iff & Y_C Y < \frac{X_C Y_\beta}{X_\beta} Y \text{ by multiplication by } Y < 0 \\ \iff & Y_C Y < -\frac{X_\beta f(X) X_C}{X_\beta} \text{ by (6.28)} \\ \iff & Y_C Y < -X_C f(X). \end{aligned}$$

This last inequality is exactly $\partial_C E < 0$, and the proof is complete. □

**Remark 6.4.** *Note that we did not use $E = 0$ to prove $\partial_C E < 0$. Therefore, our proof applies for any $t$: the derivative of $E$ with respect to $C$ is negative at the point where $E$ is minimal (with respect to the initial data $\beta$). However, we only use this property when the minimum of $E$ is equal to 0 for our purpose.*

The proposition below is equivalent to Proposition 6.4, thanks to Proposition 6.13. It describes the asymptotic behavior of the limit set as $L$ goes to 0 or to $+\infty$.

**Proposition 6.14.** *The function $C_*$ is non-increasing and satisfies*

*(i)* $\lim_{L\to\infty} C_*(L) = c_*(f)$,

*(ii)* $C_*(L) \sim \frac{1}{4L} \log\left(1 - \frac{F(1)}{F(\theta)}\right)$ *when* $L \to 0$.

*Proof.* The proof of (ii) is a direct consequence of Lemma 6.6. Indeed from estimate (6.26) we deduce that $\lambda$ goes to 0 only if $\beta - \alpha \to 0$. It can occur only if $\beta < \theta_c$. Then with (6.25), we deduce that when $L \to 0$, we have

$$C_*(L) \sim \frac{1}{4L} \min_\beta \ln\left(1 - \frac{F(1)}{F(\beta)}\right) = \frac{1}{4L} \ln\left(1 - \frac{F(1)}{F(\theta)}\right).$$

For the point (i), we have by Proposition 6.11 that for all $\epsilon > 0$, there exists $\alpha_\epsilon$ (close to 0) and $\beta_\epsilon$ (close to 1) such that

$$\gamma(\alpha_\epsilon, \beta_\epsilon) = c_*(f) + \epsilon.$$

Simultaneously, $\lambda(\alpha_\epsilon, \beta_\epsilon) \to +\infty$ as $\epsilon \to 0$. Thus $\lim_{L\to+\infty} C_*(L) = c_*(f)$.

□

We now state two auxiliary facts before getting to the proof of our last main result (remaining parts of Proposition 6.3):

**Proposition 6.15.** *For all $C \geq c_*(f)$ there exists unique $\alpha_C$ and $\beta_C$ such that the generalized problem (6.13) (i.e. we impose that its solutions are of class $\mathcal{C}^1$ and let $L = +\infty$) has solutions with $(\alpha, \beta) = (\alpha_C, 1)$ and $(\alpha, \beta) = (0, \beta_C)$. When $C = c_*(f)$ this property holds with $(\alpha, \beta) = (0, 1)$: $\alpha_{c_*(f)} = 0$ and $\beta_{c_*(f)} = 1$ for the (unique) traveling wave.*

*The functions $C \mapsto \alpha_C$ and $C \mapsto \beta_C$ are respectively increasing and decreasing. They converge to 0 and 1, respectively, as $C \to +\infty$*

*Conversely, for any $\alpha \in [0, \theta_c)$ there exists a unique $C \geq c_*(f)$ such that $\alpha = \alpha_C$. For any $\beta \in (0, 1]$, there exists a unique $C \geq c_*(f)$ such that $\beta = \beta_C$.*

*Proof.* First we introduce, for all $\alpha \in (0, \theta_c)$ and $\beta \in (0, 1)$:

$$C_\alpha := \lim_{\beta\to 1} \gamma(\alpha, \beta), \quad C^\beta := \lim_{\alpha\to 0} \gamma(\alpha, \beta).$$

Let us fix $C > c*(f)$. We are going to show that there exists a unique $\alpha \in (0, \theta_c)$ such that $C_\alpha = C$. To this aim, we notice that $\alpha \mapsto C_\alpha$ is continuous, increasing (from Proposition 6.9) and

$C_0 = c_*(f)$. Then it suffices to prove that $\lim_{\alpha\to\theta_c} C_\alpha = +\infty$. Once this will be done, defining $\alpha_C$ by $C_{\alpha_C} = C$ will yield the result.

Similarly, we are going show that there exists a unique $\beta \in (0,1)$ such that $C^\beta = C$. Again, we notice that $\beta \mapsto C^\beta$ is continuous, decreasing, and $C^1 = c_*(f)$. Then it suffices to prove that $\lim_{\beta\to 0} C^\beta = +\infty$.

Let $C_{\theta_c} := \lim_{\alpha\to\theta_c} C_\alpha, \quad C^0 := \lim_{\beta\to 0} C^\beta$. We are going to prove $C_{\theta_c} = C^0 = +\infty$.

The claim for $C^0$ is a straightforward consequence of Proposition 6.10. For $C_{\theta_c}$, let us assume by contradiction that $C_{\theta_c} < +\infty$. In this case we find a solution to

$$\begin{cases} -p'' - C_{\theta_c} p' = f(p) \text{ on } (-\infty, 0) \\ -p'' = f(p) \text{ on } (0, +\infty), \\ p(-\infty) = 1,\ p(+\infty) = 0, \end{cases} \tag{6.31}$$

such that $p(0) = \theta_c$. Multiplying the equation by $p'$ and integrating over $(0, +\infty)$ yields $p'(0) = 0$. However, this cannot hold because by hypothesis ($f$ is bistable), $f(\theta_c) > 0$, and then this imposes $p''(0) < 0$: $p$ would reach a local maximum at 0, which contradicts the fact that it has to decrease on $(-\infty, 0)$. (Similarly, Hopf Lemma gives that $p'(0) < 0$, which contradicts $p'(0) = 0$.) □

**Remark 6.5.** *In other words, $\alpha_C$ and $\beta_C$ may be defined respectively as $\alpha_C = p(0)$ where $p$ is the unique solution of class $\mathcal{C}^1$ of*

$$\begin{cases} -p'' - Cp' = f(p) \text{ on } (-\infty, 0), \\ -p'' = f(p) \text{ on } (0, +\infty), \\ p(-\infty) = 1,\ p(+\infty) = 0,\ p > 0. \end{cases}$$

*and as $\beta_C = p(0)$ where $p$ be the unique solution of class $\mathcal{C}^1$ of*

$$\begin{cases} -p'' = f(p) \text{ on } (-\infty, 0), \\ -p'' - Cp' = f(p) \text{ on } (0, +\infty), \\ p(-\infty) = 1,\ p(+\infty) = 0,\ p > 0. \end{cases}$$

*See [99] for existence and uniqueness of these solutions: the results therein apply directly up to transforming $p(\,\cdot\,)$ into $p(-\,\cdot\,)$ for the first problem, and into $1 - p(\,\cdot\,)$ for the second one.*

**Lemma 6.7.** *Let $C > c_*(f)$. For all $\beta \in (\beta_C, 1)$, there exists a unique $\alpha_C^+(\beta) \in (0, \alpha_C)$ such that $\gamma(\alpha_C^+(\beta), \beta) = C$. We introduce $L^C(\beta) := \lambda(\alpha_C^+(\beta), \beta)$.*

*At the limits, $\alpha_C^+(\beta_C) = 0$ and $\alpha_C^+(1) = \alpha_C$. In addition,*

$$\exists \lim_{\beta\to\beta_C} L^C(\beta) = \lim_{\beta\to 1} L^C(\beta) = +\infty.$$

*Hence we can define*

$$L_m(C) := \min_{\beta\in(\beta_C,1)} L^C(\beta).$$

*Then, $L_m$ is decreasing and $\lim_{C\to+\infty} L_m(C) = 0$.*

*Proof.* Existence and uniqueness for $\alpha_C^+$ (whence the definition of $L^C$) comes from the fact that the equation's flow is strictly inward on the level sets of $E$ (by Lemma 6.4).

The two limits at $\beta_C$ and 1 of $\alpha_C^+$ are straightforward, as well as those of $L^C$ (this may be seen as a corollary of Proposition 6.15). This justifies the existence of a minimum for $L^C$.

Everything being monotone with respect to $C$, this implies that $L_m$ is decreasing. Finally, the minimality of $L_m$ implies that $L_m \to 0$ as $C \to +\infty$, because (by Proposition 6.13) for all $L > 0$, there exists $C_*(L)$, $(C_*(L), L) \in \mathcal{B}(f)$. Hence, for $C \geq C_*(L)$, necessarily $L_m(C) < L$.

□

We end this subsection by stating and proving an auxiliary fact on the "limit" barrier (with minimal length, equal to $L_*(C)$, at a fixed logarithmic gradient $C$). This fact is not directly useful for proving results of Section 6.2 but receives a relevant interpretation for the biological problem in Appendix 6.6.3.

**Lemma 6.8.** *Let $C > c_*(f)$. Let $\alpha_*(C), \beta_*(C)$ be such that*

$$\gamma\big(\alpha_*(C), \beta_*(C)\big) = C, \quad 2\lambda\big(\alpha_*(C), \beta_*(C)\big) = L_*(C).$$

*Then $\alpha_*$ and $\beta_*$ have a limit as $C \to +\infty$, and*

$$\lim_{C\to\infty} \alpha_*(C) = \theta = \lim_{C\to\infty} \beta_*(C).$$

*In addition, for all $C > c_*(f)$, $\alpha_*(C) < \theta < \beta_*(C)$, and*

$$\beta_*(C) - \alpha_*(C) = \frac{1}{C}\big(\sqrt{2(F(1) - F(\theta))} - \sqrt{-2F(\theta)}\big) + o(\frac{1}{C}).$$

*Proof.* For $C > c_*(f)$, there exists $p = p_*^C$ a solution (recall that it is not necessarily unique) of

$$\begin{cases} -p'' - Cp' = f(p), \\ \frac{1}{2}p'(-L_*(C))^2 + F(p(-L_*(C))) = F(1), \quad \frac{1}{2}p'(L_*(C))^2 + F(p(L_*(C))) = 0. \end{cases}$$

such that

$$p_*^C\big(L_*(C)\big) = \alpha_*(C), \quad p_*^C\big(-L_*(C)\big) = \beta_*(C).$$

We define $v_C : [-1, 1] \to [0, 1]$ by $v_C(x) = p_*^C(xL_*(C))$. Then $v_C$ satisfies

$$\begin{cases} -v_C'' - CL_*(C)v_C' = \big(L_*(C)\big)^2 f(v_C) \\ \dfrac{1}{2\big(L_*(C)\big)^2}v_C'(-1)^2 + F(v_C(-1)) = F(1), \quad \dfrac{1}{2\big(L_*(C)\big)^2}v_C'(1)^2 + F(v_C(1)) = 0. \end{cases}$$

We introduce $y = v_C'$. Recalling that $CL_*(C) \sim_{C\to\infty} \frac{1}{4}\log(1 - \frac{F(1)}{F(\theta)})$ (by Proposition 6.14), for all $z \in (-1, 1)$ we find

$$y(z) = y(-1)e^{-CL_*(C)(z+1)} + O(\frac{1}{C^2}).$$

It follows that $v_C(z) = v_C(-1) + \frac{v_C'(-1)}{CL_*(C)}\big(1 - e^{-CL_*(C)(z+1)}\big) + O(\frac{1}{C^2})$.

Hence $v_C(1) = v_C(-1) + \frac{v_C'(-1)}{CL_*(C)}\big(1 - e^{-2CL_*(C)}\big) + O(\frac{1}{C^2})$ and $v_C'(1) = v_C'(-1)e^{-2CL_*(C)} + O(\frac{1}{C^2})$. From this, we deduce

$$\begin{cases} \frac{1}{2\big(L_*(C)\big)^2}v_C'(-1)^2 + F(v_C(-1)) = F(1) \\ \frac{1}{2\big(L_*(C)\big)^2}v_C'(-1)^2 e^{-4CL_*(C)} + F\Big(v_C(-1) + \frac{v_C'(-1)}{CL_*(C)}\big(1 - e^{-2CL_*(C)}\big)\Big) = O(\frac{1}{C^2}) \end{cases} \tag{6.32}$$

Let $z = v_C(-1)$ and $y = v_C'(-1)$. The first equation gives $y = O(1/C)$, so at the limit $C \to \infty$ we find $\lim_{C\to\infty} v_C(-1) = \lim_{C\to\infty} v_C(1)$: $v_C$ itself converges to a constant $z_\infty$. Using the first equation in the second we find

$$(F(1) - F(z))e^{-4CL_*(C)} + F\big(z + O(\frac{1}{C})\big) = O(\frac{1}{C^2}).$$

Recalling that $e^{4CL_*(C)} \xrightarrow[C\to\infty]{} 1 - \frac{F(1)}{F(\theta)}$ we recover as $C \to \infty$

$$F(1) - F(z_\infty) + (1 - \frac{F(1)}{F(\theta)})F(z_\infty) = 0,$$

that is $F(1)\big(1 - \frac{F(z_\infty)}{F(\theta)}\big) = 0$, or equivalently $F(z_\infty) = F(\theta)$.

Hence $\lim_{C\to\infty} z = \theta$. Recalling $z = v_C(-1) = \beta_*(C)$, we find that both $\alpha_*(C)$ and $\beta_*(C)$ converge to $\theta$.

Let us fix $C > c_*(f)$. For all $\alpha \in (0, \alpha_C)$, there exists a unique $\beta(C, \alpha)$ such that $\gamma(\alpha, \beta(C, \alpha)) = C$. Obviously, $\alpha \mapsto \beta(C, \alpha)$ is increasing.

Then, we claim that if $\theta \leq \alpha_0 < \alpha_1 < \alpha_C$ then $\lambda(\alpha_0, \beta(C, \alpha_0)) < \lambda(\alpha_1, \beta(C, \alpha_1))$. Symmetrically, if $\alpha_0 < \alpha_1 < \alpha_C$ are such that $\beta(C, \alpha_1) < \theta$, then $\lambda(\alpha_0, \beta(C, \alpha_0)) > \lambda(\alpha_1, \beta(C, \alpha_1))$. This is a simple consequence of the expression of $\lambda$ and of the fact that $F$ is decreasing on $[0, \theta]$, increasing on $[\theta, 1]$.

Differentiating (6.16) with respect to $p$, choosing $\alpha = \alpha_*(C)$ and $\beta = \beta_*(C)$ and integrating between $\alpha$ and $\beta$ yields

$$C\sqrt{2(w-F)(p)} = C\sqrt{-2F(\alpha)} + C^2(p-\alpha) - C\int_\alpha^p \frac{f(p')dp'}{\sqrt{2(w-F)(p')}}.$$

From this we get

$$2CL_*(C) = C\int_\alpha^\beta \frac{dp}{\sqrt{2(w-F)(p)}} = \int_\alpha^\beta \frac{dp}{p - \alpha + \frac{\sqrt{-2F(\alpha)}}{C} - \frac{1}{C}\int_\alpha^p \frac{f(p')dp'}{\sqrt{2(w-F)(p')}}}. \tag{6.33}$$

By Proposition 6.14 we know that $2CL_*(C) = \frac{1}{2}\log\left(\frac{F(1)-F(\theta)}{-F(\theta)}\right) + o(1)$ (where the $o$ is taken as $C \to \infty$). Rewriting the right-hand side of (6.33) (recalling that $\beta_* - \alpha_* = o(1)$), we find

$$\frac{1}{2}\log\left(\frac{F(1)-F(\theta)}{-F(\theta)}\right) = \log\left(1 + C\frac{\beta-\alpha}{\sqrt{-2F(\alpha)}}\right) + o(1).$$

Since $\alpha \to \theta$ as $C \to +\infty$, taking the exponential of both sides we obtain

$$(1+o(1))\sqrt{2(F(1)-F(\theta))} = \sqrt{-2F(\theta)} + C(\beta_*(C) - \alpha_*(C)),$$

and the claim is proved.

□

### 6.5.6 Gathering the results on the barrier set.

We can now prove the remaining parts of Proposition 6.3, concerning order and extremal elements (recalling the first point has been stated and proved in Lemma 6.1).

*Proposition 6.3.* First, we know the $\alpha$'s and the $\beta$'s are in the same order. More precisely, if there are $(C, L)$-barriers from $\beta_0$ to $\alpha_0$ and from $\beta_1$ to $\alpha_1$, and $\beta_0 < \beta_1$, then $\alpha_0 < \alpha_1$ by Proposition 6.12. We then crucially use Lemma 6.5.

Applying Lemma 6.5 to two barriers, on $[-L, L]$ (or equivalently on $[0, 2L]$, to fit the notations in (6.19)) yields the global ordering of all barriers. Barriers obviously satisfy $X > 0$, $Y < 0$, by Lemma 6.1

Now we take $\lambda_+$ associated with maximal $\beta_+ = p_{\lambda_+}(-L)$ and $\alpha_+ = p_{\lambda_+}(L)$. For all $\epsilon > 0$ small enough, we construct a subsolution to (6.6) by letting

$$\begin{cases} -p_\epsilon'' - Cp_\epsilon' = f(p_\epsilon) \text{ in } (-L, L),\ p_\epsilon(-L) = \beta_+ + \epsilon, \\ -p_\epsilon'' = f(p_\epsilon) \text{ in } \mathbb{R} - (-L, L), \\ F(p_\epsilon(-L)) + \frac{1}{2}(p_\epsilon'(-L))^2 = F(1),\ F(p_\epsilon(L)) + \frac{1}{2}(p_\epsilon'(L^+))^2 = 0 \end{cases} \tag{6.34}$$

where $p_\epsilon$ is continuous, but $p_\epsilon'$ exhibits a jump at $L$.

Then we can prove that $p_\epsilon(L) > p_{\lambda_+}(L)$ and the jump has the good sign to provide a subsolution $p_\epsilon'(L^-) < p_\epsilon'(L_+)$, by maximality of $\beta_+$. The second point can be seen easily in the phase plane. It is in fact a straightforward consequence of the continuity of $\beta \mapsto E(2L; C, \beta)$

Now, it remains to see that $p_\epsilon(x) > p_{\lambda_+}(x)$ for all $x \in [-L, L]$, hence for all $x \in \mathbb{R}$. In fact, this is a simple consequence of Lemma 6.5. One simply has to check that by continuity of the solutions of differential equations with respect to the initial data, for $\epsilon > 0$ small enough, $p_\epsilon$ remains in $(0, 1)$ on $[-L, L]$ and $p_{\epsilon'}$ remains negative.

The proof is totally similar for the stability from below of $p_{\lambda_-}$ (defined by minimality of $\beta_- = p_{\lambda_-}(-L)$ and $\alpha_- = p_{\lambda_-}(L)$, making use of Lemma 6.5 again, hence we don't reproduce it here.

The last point comes from the fact that $\lambda(\alpha_C^+(\beta), \beta)$, which is defined on $(\beta_C, 1)$, goes to $+\infty$ at $\beta_C$ and at 1 (Lemma 6.7), hence reaches its minimum (which is necessarily equal to $L_*(C)$) at some $\beta_0(C) \in (\beta_C, 1)$. For $L > L_*(C)$, there exists $(\beta_1, \beta_2)$ with $\beta_C < \beta_1 < \beta_0(C)$ and $\beta_0(C) < \beta_2 < 1$ such that $\lambda(\alpha_C(\beta_1), \beta_1) = \lambda(\alpha_C(\beta_2), \beta_2) = L$, yielding two distinct barriers defined by $(\alpha_C(\beta_i), \beta_i)$ for $i \in \{1, 2\}$. □

**Remark 6.6.** *We interpret Proposition 6.3 in terms of asymptotic behavior of solutions to (6.4) thanks to Proposition 6.2. Any initial datum below $p_{\lambda_-}$ will be unable to pass and propagate (the wave it may have "initiated" on $(-\infty, -L)$ will be blocked), while any initial datum above $p_{\lambda_+}$ will propagate.*

**Remark 6.7.** *Proposition 6.3 applies in particular when there exists a unique $(C, L)$ barrier (which should generically hold when $L = L_*(C)$). In this case, this barrier is simultaneously stable from below and unstable from above. As before, either the solution is blocked below this barrier ("stable from below"), or the solution passes the barrier, in which case it propagates to $+\infty$ ("unstable from above").*

### 6.5.7 Generalizing the barriers.

Now we move to the proof of Corollary 6.1.

**Remark 6.8.** *If $\mathcal{Y} := \{C\mathbb{1}_{[-L,L]},\ C, L > 0\}$, then $\mathcal{B}(f) = \mathcal{B}_{\mathcal{Y}}(f)$ (in fact, (6.6) is a special case of (6.9)).*

First, we note that these "generalized" barriers are still decreasing, as long as $\eta$ is. The set of gradient profiles $\mathcal{X}$ was introduced in (6.8)

**Lemma 6.9.** *For $\eta \in \mathcal{X}$, a $\eta$-barrier is necessarily monotone decreasing.*

*Proof.* Let $L > 0$ be such that $Supp(\eta) \subseteq [-L, L]$.

For $x \in (-\infty, -L)$, since $-p'' = f(p)$ we get by multiplication by $p'$ and integration:

$$\frac{1}{2}(p'(x))^2 + F(p(x)) = F(1).$$

Hence $p'$ cannot vanish unless $p = 1$, which is impossible.

Now, for $x \in (L, +\infty)$ we get similarly

$$\frac{1}{2}(p'(x))^2 + F(p(x)) = 0,$$

so $p'$ can vanish only if $p = 0$ or $p = \theta_c$. As before, $p = 0$ is impossible. We will show that $p(L) < \theta_c$, which is equivalent to $p'(L) \neq 0$, and will be done.

For $x \in (-L, L)$, we define $E(x) := \frac{1}{2}(p'(x))^2 + F(p(x))$. Then

$$E'(x) = -\eta(x)p'(x)^2 \leq 0,$$

so $E$ is non-increasing. (Here it is crucial that $\eta \in \mathcal{X} \implies \eta \geq 0$.) In addition, $E(-L) = F(1)$ and $E(L) = 0$.

Let $x_m := \inf\{x > -L,\ p'(x) = 0\}$ and assume by contradiction $x_m \leq L$. If $p(x_m) < \theta_c$ then $E(x_m) = 0 + F(p(x_m)) < 0$, which is absurd because $E$ is non-increasing and $E(L) = 0$. We are left with $p(x_m) \geq \theta_c > \theta$. This implies that $p''(x_m) = -f(p(x_m)) - \eta(x_m)p'(x_m) < 0$. In this case, $p$ reaches a local maximum at $x_m$, which is absurd because by definition of $x_m$, $p' < 0$ on $(-L, x_m)$.

Hence $p$ is monotone decreasing. □

**Proposition 6.16.** *For all $\eta, \eta_1 \in \mathcal{X}$, $\eta \in \mathcal{B}_{\mathcal{X}}(f) \implies \eta + \eta_1 \in \mathcal{B}_{\mathcal{X}}(f)$.*

*If $\lambda > 0$ then $\eta \in \mathcal{B}_{\mathcal{X}}(f)$ is equivalent to $\lambda\eta(\lambda\,\cdot) \in \mathcal{B}_{\mathcal{X}}(\lambda^2 f)$. This point enables us to assume $F(1) = 1$ without loss of generality.*

*Proof.* The last two points are simple: apart from $\eta$ the rest of the problem is translation-invariant; $q(x) := p(\lambda x)$ satisfies

$$-\frac{1}{\lambda^2}q''(x) - \frac{1}{\lambda}\eta(\lambda x)q'(x) = f(q(x)) \text{ on } \mathbb{R}.$$

Multiplying this equation by $\lambda^2$ yields the result.

The first point however requires a complete proof, which mimics that of Proposition 6.5. Let $p_\eta$ be a $\eta$-barrier. Then

$$-p_\eta'' - (\eta + \eta_1)p_\eta' \geq -p_\eta'' - \eta p_\eta' = f(p_\eta).$$

Hence $p_\eta$ is a super-solution to the $(\eta + \eta_1)$-problem.

Simultaneously, as in the proof of Proposition 6.5, the (translated) $\alpha$-bubble gives a sub-solution to the $(\eta + \eta_1)$-problem which lies below $p_\eta$.

By the sub- and super-solution method, this provides a $(\eta + \eta_1)$-barrier. □

Then, Corollary 6.1 follows directly from the first point (positivity) in Proposition 6.16 and Theorem 6.2.

## 6.6 Discussion and extensions

### 6.6.1 Summary of the results

Before discussing the derivation of the models and some extensions of our results, we sum up the content of the article.

On the first hand, thanks to a change of variables, we established a sharp threshold property for equation (6.3) in the bistable case and gave a full description of the situation in the KPP case (Theorem 6.1). Therefore in this simple and homogeneous model, when total population is approximated as a function of infection frequency, no stable propagation blocking can occur. We also described the propagules in this case (Proposition 6.1).

On the other hand, when the total population is increasing along a line, we characterized the constant logarithmic gradients that create stable blocking fronts (Theorem 6.2), and gave a sufficient condition in Corollary 6.1 for the non-constant case. We stated the asymptotic behavior of solutions in Proposition 6.2, when there are no barriers or when initial data can be compared to some of the barriers. Then, a deeper understanding of the barriers (Proposition 6.3) and of the barrier set (Proposition 6.4) enabled us to describe the important "unstable front" associated with stable blocking fronts. Computing this unstable front in the context of a blocked artificial introduction of *Wolbachia*, for example, may help designing future releases of infected mosquitoes in order to clear the propagation hindrance.

The remainder of this section is organized as follows. We explain in Subsection 6.6.2 how (6.3) and (6.2) are derived from a two-population model, then in Subsection 6.6.3 we discuss the link between the barriers we considered in this paper and the local barrier studied in [29], and finally we gather in Subsection 6.6.4 some numerical conjectures we were not able to prove so far.

### 6.6.2 Derivation from a two-population model

Both (6.3) and (6.2) may be derived in some sense from a single two-population model (posed on $\mathbb{R}^d$ with $d \in \{1, 2\}$). To perform this derivation we consider the model for infected and uninfected mosquitoes proposed in [211]. We denote by $n_i$, resp $n_u$, the density of infected, resp. uninfected, mosquitoes.

$$\partial_t n_i - D\Delta n_i = (1 - s_f)F_u n_i\big(1 - \frac{N}{K}\big) - \delta d_u n_i, \tag{6.35}$$

$$\partial_t n_u - D\Delta n_u = F_u n_u(1 - s_h p)\big(1 - \frac{N}{K}\big) - d_u n_u. \tag{6.36}$$

This model uses 7 parameters: $D$ is the (constant) diffusion rate, $F_u$ is the fecundity of uninfected mosquitoes, $s_f \in (0, 1)$ is a dimensionless parameter taking into account the fecundity reduction for infected mosquitoes ($F_i = (1 - s_f)F_u$ is the fecundity of infected mosquitoes), $K$ is the environmental capacity, $d_u$ is the death rate, $d_i = \delta d_u$ is the death rate of infected mosquitoes ($\delta > 1$), $s_h \in (0, 1)$ is the cytoplasmic incompatibility parameter.

We introduce the total population $N = n_i + n_u$ and the fraction of infected mosquitoes $p = \frac{n_i}{n_i+n_u}$. After straightforward computations, we obtain the system

$$\partial_t N - D\Delta N = N\left(F_u\big(1-\frac{N}{K}\big)\big((1-s_f)p+(1-p)(1-s_h p)\big) - d_u(\delta p + 1 - p)\right), \tag{6.37}$$

$$\partial_t p - D\Delta p - 2D\frac{\nabla p\cdot\nabla N}{N} = p(1-p)\left(F_u\big(1-\frac{N}{K}\big)(s_h p - s_f) + d_u(1-\delta)\right). \tag{6.38}$$

We make the assumption of large fecundity (as in [211]) and introduce $\epsilon \ll 1$, so that $F_u$ scales as $\widetilde{F}_u/\epsilon$, and we can rewrite (6.37) as

$$\partial_t N - D\Delta N = N\left(\widetilde{F}_u\big(\frac{1}{\epsilon}-\frac{N}{K}\big)\big((1-s_f)p+(1-p)(1-s_h p)\big) - d_u(\delta p + 1 - p)\right). \tag{6.39}$$

Assuming in this way that fecundity is of a bigger order of magnitude than the death rate appeared as a technical assumption in [211] to recover a proper limit as $\epsilon$ goes to 0 for the equation on the infected proportion $p$. Bio-ecology of *Aedes* mosquitoes gives a quick but relevant justification of this assumption by the process of "skip oviposition": the availability of good-quality containers affects the egg-laying behavior of females, inducing more extensive and energy-consuming search when breeding sites are scarce. This phenomenon has been documented in [48] (for *Ae. aegypti*) and [62] (for *Ae. albopictus*), for example.

There is more to say about the values of the parameters. In the modeling work [50], the authors used the experimental values of [238] at a temperature of $30°C$. In our model these values would yield $F_u \simeq 10$ day$^{-1}$ and $d_u \simeq 0.3$ day$^{-1}$ (taking into account the immature stages mortality). It makes sense to consider that the dimensionless quantity $F_u/d_u \simeq 33$ is large. Estimating the diffusion coefficient may be difficult, but here we are mainly interested in the relative orders of magnitude of the parameters. Sticking to the choices of [50] we can consider $D \simeq 1.25\times 10^{-2}$ km$^2$.day$^{-1}$.

We also assume that the carrying capacity may depend on space, that is $K = K(x)$. Then we introduce a development of $N = N^\epsilon(t,x)$ by letting

$$N = N^\epsilon(t,x) = K(x)\Big(1-\epsilon z^\epsilon(t,x) + \epsilon^2 w^\epsilon(t,x)\Big).$$

Equating the leading terms in (6.39) yields

$$z^\epsilon(t,x) = \frac{d_u\big((\delta-1)p(t,x)+1\big) - D\Delta K(x)/K(x)}{(1-s_f)p(t,x)+(1-p(t,x))(1-s_h p(t,x))}.$$

We know that $z^\epsilon$ is uniformly bounded with respect to $\epsilon$. Under the assumption that for all $x\in\mathbb{R}^d$, $d_u\min(1,\delta) > D\Delta K(x)/K(x)$ we claim that $w^\epsilon$ is uniformly bounded with respect to $\epsilon$.

Equation (6.38) then rewrites

$$\partial_t p - D\Delta p - 2D\frac{\nabla K}{K}\cdot\nabla p - 2D\nabla p\cdot\nabla\big(\log(1-\epsilon z^\epsilon+\epsilon^2 w^\epsilon\big) = \\ p(1-p)\Big((s_h p - s_f)\frac{d_u\big((\delta-1)p+1\big) - D\Delta K/K}{(1-s_f)p+(1-p)(1-s_h p)} - d_u(\delta-1)\Big). \tag{6.40}$$

We study (6.40) with two different settings. First, we assume that $K$ is constant. Like in Section 6.4.1, let $h_\epsilon(p) = 1 - \epsilon\frac{d_u}{\sigma F_u}h_0$. In this case (6.40) rewrites

$$\partial_t p - D\Delta p + 2|\nabla p|^2\frac{h'_\epsilon(p)}{h_\epsilon(p)} = p(1-p)\Big(d_u(s_h p - s_f)h(p) - d_u(\delta-1) - \epsilon w\Big). \tag{6.41}$$

Neglecting the $\epsilon w^\epsilon$ term in the right-hand side of (6.41) yields problem (6.3), with $h'_\epsilon/h_\epsilon$ as a first-order correction term. We claim that this approximation does not change the structure of the problem because the right-hand side always keeps a bistable structure, for $\epsilon$ small enough.

In a second setting we assume that $K$ varies with space and keep only the leading terms to obtain

$$\partial_t p - D\Delta p - 2D\frac{\nabla K}{K}\cdot\nabla p = p(1-p)\Big((s_h p - s_f)\frac{d_u\big((\delta-1)p+1\big) - D\Delta K/K}{(1-s_f)p+(1-p)(1-s_h p)} - d_u(\delta-1)\Big). \tag{6.42}$$

Then, we get problem (6.2) from (6.42) under the assumption that $D\Delta K/K$ is much smaller than $d_u \min(1,\delta)$. In order to understand the actual meaning of this assumption, let $\delta \geq 1$. If $d = 1$ and $K(x) = K_0 e^{\frac{C}{2}(x-x_0)}$ on some set $(x_0, x_1) \subset \mathbb{R}$ and is constant otherwise (this is the specific case we studied in this article, see (6.5)), then $\Delta K/K$ is either equal to 0 or to $C^2/4$, and $C$ can be expressed for example in $\text{km}^{-1}$. On the area $(x_0, x_1)$ where carrying capacity varies, the population is doubled every $2\log(2)/C$ kilometer. Our assumption $D\Delta K/K < d_u$ with the values of [238] at $30°C$ then reads approximately $C < 2\sqrt{0.3/0.0125} \simeq 9.8$ $\text{km}^{-1}$, which means that the carrying capacity is at most doubled approximately every 140 m. If the gradient is steeper than this, then our approximation does not hold in $(x_0, x_1)$. All in all, for $C \in \left(\frac{c_*}{D}, 2\sqrt{\frac{d_u}{D}}\right)$, where $c_*$ is the (unique) bistable traveling wave speed (which scales by the way as $\sqrt{d_u D}$), we get a gradient value which allows for wave blocking (if applied on a large enough area) while justifying our approximation. With these parameter values, $c_*$ is approximately equal to 7.2 $\text{m.day}^{-1}$, so that the range for $C$ is approximately equal to $(1.59, 9.8)$ (in $\text{km}^{-1}$). Here, we chose $\delta = 10/9$, $s_f = 0.1$ and $s_h = 0.8$. Note that taking different values of $\delta, s_f$ and $s_h$ would yield a different value of $c_*$ (independently from $d_u$ and $D$), possibly allowing a wider range for $C$.

Elaborating on this derivation suggests the study of another problem of interest, which would complement the simplification (6.2) (inspired by the seminal work [29]) studied in the present article. It consists of (6.42), where we keep the term in $\Delta K/K$ rather than neglecting it. This is a direction for future works on this topic.

### 6.6.3 Critical population jump

In this section we make a link with the concept of barrier strength used in [29] for local barriers. First, we define

**Definition 6.3.** *A **local barrier** is a jump (i.e. a discontinuity) in the size of the total population $N$ which is sufficient to block a propagating wave.*

Starting from our $(L, C)$-barriers, we get a local barrier by letting $L \to 0$. Simultaneously, we scale $C$ as $C(\alpha(L), \beta(L); L)$ for some $\alpha(L) < \beta(L)$. The jump in the total population, from $N_L$ (on the left) to $N_R > N_L$ (on the right) always reads

$$N_R = \exp\Big(\int_{-L}^{L} \frac{C}{2} dx\Big) N_L = \exp(LC) N_L.$$

The limit equation as $L \to 0$ reads

$$\begin{cases} -p'' - \lim_{L\to 0} \left\{C(\alpha(L), \beta(L); L) \mathbb{1}_{-L\leq x\leq L} p'\right\} = f(p) & \text{on } \mathbb{R}, \\ p(0^-) = \beta_0, \quad p(0^+) = \alpha_0, \\ p(-\infty) = 1, \quad p(+\infty) = 0, \end{cases} \tag{6.43}$$

where we assumed $\alpha(L) \xrightarrow[L\to 0]{} \alpha_0$, $\beta(L) \xrightarrow[L\to 0]{} \beta_0$. Now, recall that by (6.26), necessarily $\alpha_0 = \beta_0$.

This means that $N = N_L$ on $(-\infty, 0)$ and $N = N_R$ on $(0, +\infty)$, with

$$N_R = e^{K(\alpha_0)} N_L,$$

where $K(\alpha_0) = \lim_{L\to 0} L \cdot C(\alpha(L), \beta(L); L)$. $K$ depends only on $\alpha_0$ indeed: by formula (6.25) in Lemma 6.6,

$$K(\alpha_0) = \frac{1}{4} \log\Big(1 - \frac{F(1)}{F(\alpha_0)}\Big).$$

This implies that

$$N_R = \Big(1 - \frac{F(1)}{F(\alpha_0)}\Big)^{1/4} N_L.$$

Equation (6.43) then rewrites

$$\begin{cases} -p'' + \frac{1}{4} \log\left(1 - \frac{F(1)}{F(\alpha_0)}\right) \langle \delta_0', p\rangle = f(p) & \text{on } \mathbb{R}, \\ p(0) = \alpha_0, \\ p(-\infty) = 1, \quad p(+\infty) = 0, \end{cases} \tag{6.44}$$

and the derivation of (6.44) is legitimate for $\alpha_0 = \lim_{L\to 0} \alpha(L) = \theta$, by Lemma 6.8.

As a consequence,

**Proposition 6.17.** *The minimal "jump" in the total population that can block a wave is:*

$$N_R = \left(1 - \frac{F(1)}{F(\theta)}\right)^{1/4} N_L.$$

If we understand [29] correctly, the authors addressed the situation where for (6.43), $p'(0^-) = p'(0^+)$. In view of our result, it means $F(1) = 0$. But simultaneously they wanted $p(0^-) \neq p(0^+)$. We find that this cannot be obtained by using equation (6.2). However, if the reaction term $f$ depends itself on $N$ (as it is expected to do, see Section 6.6.2), then this becomes possible.

A good intuition is that the stronger the population gradient, the smaller the population "jump" required for blocking. In the limit of a real, discontinuous jump, we recover the critical value from Proposition 6.17.

We can state this result in more generality using the notations of this paper.

**Proposition 6.18.** *Let $\mathcal{H}(f,K) := \{C > c_*(f),\ (C, \frac{K}{C}) \in \mathcal{B}(f)\}$. There exists a minimal $K_0(f) > 0$ such that if $K > K_0(f)$ then $\mathcal{H}(f,K)$ is non-empty.*

*Proof.* We remark that $(C, K/C) \in \mathcal{B}(f)$ if and only if $K \geq CL_*(C)$, by Theorem 6.2.

Let $K_0 = \min_C CL_*(C) > 0$, and $K > K_0$. Then there exists at least one $C(K) > c_*(f)$ such that $C(K)L_*(C(K)) = K$.

□

Assuming $C \mapsto CL_*(C)$ is decreasing (as seems to be the case, see Figure 6.6 above), a stronger result holds, which confirms the above intuition. In this case, $\mathcal{H}(f,K)$ is equal to a half-line for any $K > \left(1 - \frac{F(1)}{F(\theta)}\right)^{1/4}$, and is empty otherwise. We refer to 6.A for further discussion on this topic.

### 6.6.4 Numerical conjectures

About Lemma 6.8, it is a numerical conjecture that for *generic* bistable function $f$, $\alpha_*$ is increasing, $\beta_*$ is decreasing, and both are uniquely defined (see Figure 6.8).

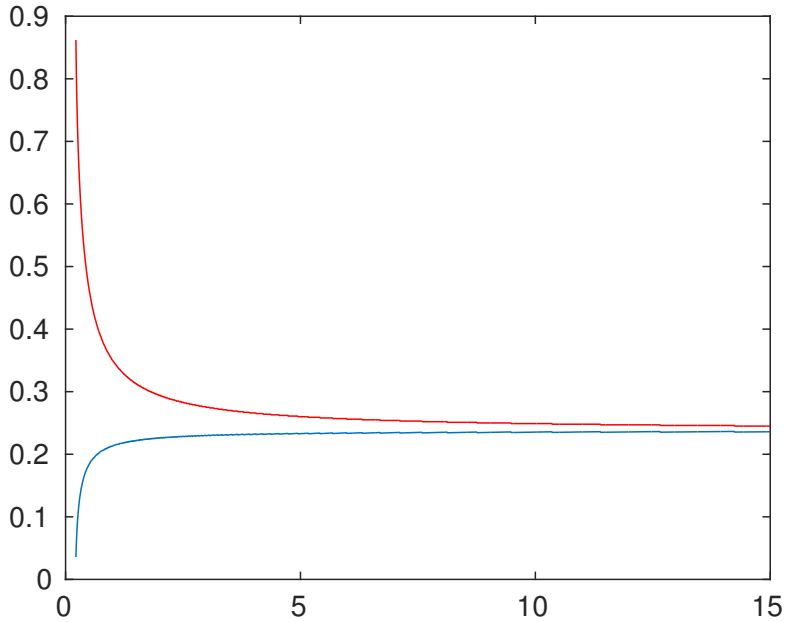



Figure 6.8: Plot of $\alpha_*$ (in blue, below) and $\beta_*$ (in red, above) as functions of $C$ (respectively increasing and decreasing), obtained by simulating the ODE system (6.19) with $f$ as in Subsection 6.2.3.

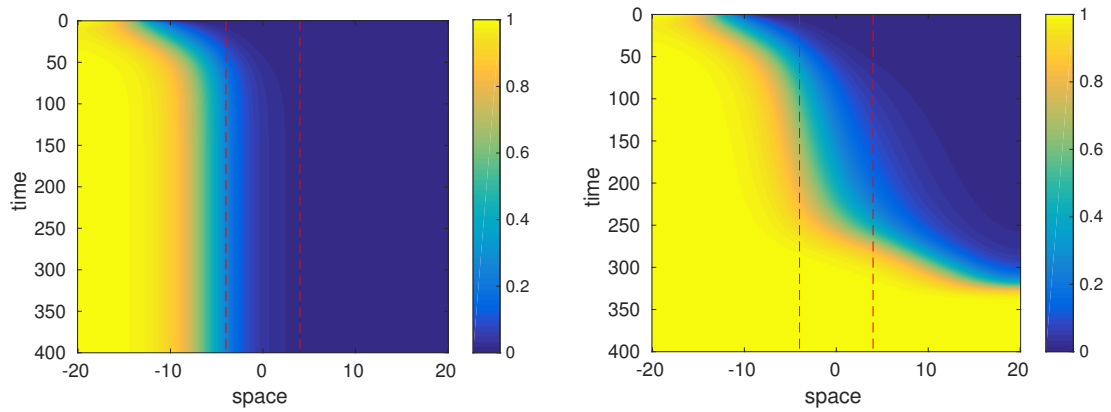



Figure 6.9: Plot of the proportion of the invading population with respect to time (y-axis) and space (x-axis). These are two numerical simulations of the two-population model (6.35)-(6.36) with same front-like initial data, $L = 4$ (space interval with non-zero carrying capacity gradient $[-L, L]$ marked by the two vertical dotted red lines) and two different carrying capacities. We recover the same behavior as for the single population model (6.2) *Left*: Blocking for $C = 0.2$. *Right*: Propagation for $C = 0.1$.

For *generic* bistable functions, we also conjecture that there exists exactly two barriers when $L > L_*(C)$.

Finally, the behavior we identified appears, numerically, to apply in the case of the two-population model (6.35)-(6.36), where we take $K = K(x)$ a heterogeneous carrying capacity. Figure 6.9 shows an example of the propagating/blocking alternative in this setting. As in Subsection 6.2.3, color represents the value of $p$, which is here equal to $n_i/(n_u + n_i)$. We fix $L = 4$ and choose carrying capacities as

$$K(x) = K_L \exp\Big(C \min\big((x + L)_+, 2L\big)\Big).$$

## Acknowledgements

This work was supported by Inria, France and CAPES, Brazil (processo 99999.007551/2015-00), in the framework of the STIC AmSud project MOSTICAW. The authors acknowledge support from the Emergence project from Mairie de Paris, Analysis and simulation of optimal shapes - application to lifesciences. GN acknowledges partial funding from the ANR project Nonlocal: ANR-14-CE25-0013 funded by the French Ministry of Research. MS and NV acknowledge partial funding from the ANR blanche project Kibord: ANR-13-BS01-0004 funded by the French Ministry of Research.

# Appendices

## 6.A An additional result

In the setting of (6.4), (6.5), when $C$ and $L$ are fixed, we define the population jump as $e^{CL}$, that is the quotient between the population size at the right edge of the heterogeneous area and the population size at the left edge.

We show in this appendix that the critical population jump is typically decreasing for large values of $C$. First, we draw an interesting consequence of this fact.

By definition, $C > c_*(f)$ belongs to $\mathcal{H}(f,K)$ if and only if $K/C > L_*(C)$, that is $CL_*(C) < K$. Assuming that $C \mapsto CL_*(C)$ is decreasing, we obtain that $\mathcal{H}(f,K)$ is a half-line $(\widehat{C}(f,K), +\infty)$. The interpretation of this fact is useful: for a given population jump $e^K$ $(K > 0)$, the exponential profiles for such a population increase that induce wave blocking are those which are steep enough (that is, with $C$ large enough). In particular, by smoothing the steepness of the population increase one always favors invasion, in this case.

Now, we compute the derivative of $\kappa : C \mapsto 2CL_*(C)$ in order to estimate it as $C \to +\infty$. By the same construction as in the proof of Proposition 6.13, we have

$$\kappa(C) = Ct^*(\beta^*(C), C),$$

where $X, Y$ solve

$$\begin{cases} X' = Y, \quad X(0,C,\beta) = \beta, \\ Y' = -CY - f(X), \quad Y(0,C,\beta) = -\sqrt{2(F(1) - F(\beta)}, \end{cases}$$

$t^*(\beta, C)$ is the first time such that

$$E\big[X(t^*(\beta,C),C,\beta), Y(t^*(\beta,C),C,\beta)\big] = 0,$$

(which is well-defined for $C$ large enough and $\beta$ well-chosen) and $\beta^*(C)$ is the value of $\beta \in (0,1)$ which minimizes $t^*(\beta, C)$. In particular

Using as in the proof of Proposition 6.13 subscripts to denote partial derivatives we have

$$\kappa'(C) = t^*(\beta^*(C), C) + Ct^*_C(\beta^*(C), C) + C\beta^*_C(C)t^*_\beta(\beta^*(C), C),$$

and in addition $t^*_\beta(\beta^*(C), C) = 0$ since $\beta^*(C)$ is a minimizer. By differentiation of $E(X,Y) = 0$ with respect to $C$ we also obtain

$$-CY^2 t^*_C(\beta^*(C), C) + \beta^*_C(C)(X_\beta f(X) + Y_\beta Y) + f(X)X_C + YY_C = 0.$$

Using the condition $\partial_\beta E = 0$, we deduce that

$$\kappa'(C) = t^*(\beta^*(C), C) + \frac{f(X)X_C + YY_C}{Y^2}. \tag{6.45}$$

For large $C$ we have $X \to \theta$ uniformly and $Y \to -\sqrt{2(F(1) - F(\theta))}$ uniformly, thanks to Lemma 6.8. Let

$$A(C) := \begin{pmatrix} 0 & 1 \\ -f'(\theta) & -C \end{pmatrix},$$

then (at least formally) $(X_C, Y_C)(t)$ is close to

$$\begin{pmatrix} \widetilde{X}_C \\ \widetilde{Y}_C \end{pmatrix}(t) := e^{A(C)t} \int_0^t e^{-A(C)t'} \begin{pmatrix} 0 \\ \sqrt{2(F(1) - F(\theta)} \end{pmatrix} dt'.$$

Using the fact that $t^*$ goes to 0 as $C$ goes to $+\infty$ and that $t^*C$ converges to a positive constant $K$ as $C$ goes to $+\infty$, we deduce (after straightforward computation and estimation of $e^{A(C)t'}$ for $t' \in [0, 2L_*(C)]$) that asymptotically (as $C$ is large),

$$\kappa'(C) \sim \frac{1}{C}\big(K - e^K(e^K - 1)\big),$$

and the numerator $K - e^K(e^K - 1)$ is negative as soon as $K > 0$. To make the computation less formal one simply need to develop asymptotically $(W, Z) := (X_C - \widetilde{X}_C, Y_C - \widetilde{Y}_C)$, which solves a linear differential equation with inhomogeneous terms of magnitude at most $O(1/C)$, whence the estimation.

Therefore, at least for large values of $C$, the population jump $\kappa : C \mapsto CL_*(C)$ is decreasing. However, proving this fact for all $C > c_*(f)$ is still an open problem, and could perhaps be obtained from (6.45).

# Chapter 7

# Uncertainty quantification for the invasion success

This chapter is a joint work with Nicolas Vauchelet and Jorge P. Zubelli. It was published as an article in Mathematical Biosciences and Engineering [212].

**Abstract.** Artificial releases of *Wolbachia*-infected *Aedes* mosquitoes have been under study in the past years for fighting vector-borne diseases such as dengue, chikungunya and zika. Several strains of this bacterium cause cytoplasmic incompatibility (CI) and can also affect their host's fecundity or lifespan, while highly reducing vector competence for the main arboviruses.

We consider and answer the following questions: 1) what should be the initial condition (*i.e.* size of the initial mosquito population) to have invasion with one mosquito release source? We note that it is hard to have an invasion in such case. 2) How many release points does one need to have sufficiently high probability of invasion? 3) What happens if one accounts for uncertainty in the release protocol (*e.g.* unequal spacing among release points)?

We build a framework based on existing reaction-diffusion models for the uncertainty quantification in this context, obtain both theoretical and numerical lower bounds for the probability of release success and give new quantitative results on the one dimensional case.

## 7.1 Introduction

In recent years, the spread of chikungunya, dengue, and zika has become a major public health issue, especially in tropical areas of the planet [3, 32]. All those diseases are caused by arboviruses whose main transmission vector is the *Aedes aegypti*. One of the most important and innovative ways of vector control is the artificial introduction of a maternally transmitted bacterium of genus *Wolbachia* in the mosquito population (see [34, 129, 232]). This process has been successfully implemented on the field (see [118]). It requires the release of *Wolbachia*-infected mosquitoes on the field and ultimately depends on the prevalence of one sub-population over the other. Other human interventions on mosquito populations may require such spatial release protocols (see [7, 8] for a review of past and current field trials for genetic mosquito population modification). Designing and optimizing these protocols remains a challenging problem for today (see [102, 227]), and may be enriched by the lessons learned from previous release experiments (see [117, 178, 240])

This article studies a spatially distributed model for the spread of *Wolbachia*-infected mosquitoes in a population and its success as far as non-extinction probabilities are concerned. We address the question of the release protocol to guarantee a high probability of invasion. More precisely, what quantity of mosquitoes need to be released to ensure invasion, if we have only one release point? What if we have multiple release points and if there is some uncertainty in the release protocol? We obtain lower bounds so as to quantify the success probability of spatial spread of the introduced population according to a mathematical model.

We define here an *ad hoc* framework for the computation of this success probability. As a totally new feature added to the previous works on this topic (see [60, 100, 101, 125, 220, 239]), it involves space variable as a key ingredient. In this paper we provide quantitative estimate and numerical results in dimension 1.

It is well accepted that stochasticity plays a significant role in biological modeling. Probabilities of introduction success have already been investigated for genes or other agents into a wild biological population. The recent work [29] makes use of reaction-diffusion PDEs to describe the biological phenomena underlying sucessful introduction as cytoplasmic analogues of the Allee effect. The infection of the mosquito population by *Wolbachia* is seen as an "alternative trait", spreading across a population having initially a homogeneous regular trait. Other recent models have been proposed either to compute the invasion speed ([50]), or get an insight into the induced time dynamics of more complex systems, including humans or pathogens (see [83, 121]). In the mosquito part, models usually feature two stable steady states: invasion (the regular trait disappears) and extinction (the alternative trait disappears). Since this phenomenon is currently being investigated as a tool to fight *Aedes* transmitted diseases, the problem of determination of thresholds for invasion in this equation is of tremendous importance.

The issue of survival probability of invading species has attracted a lot of attention by many researchers. Among such we may cite [28] and [196]. We stress, however, that this is not the direction followed in this paper. In the cited articles indeed, the basic underlying model is either a stochastic PDE or its discretization, and the uncertainty concerning the initial state is not considered.

In other words, although in a deterministic model as ours one can in principle numerically check for a specific initial configuration whether the invasion by the *Wolbachia*-infected mosquitoes will be successful or not, in practice such a specific initial condition is subject to uncertainty, and therefore the uncertainty quantification of the success probability is a natural question.

Our modeling goes as follows: We consider on a domain $\Omega \subseteq \mathbb{R}^d$ (usually $d \in \{1, 2\}$ and $\Omega = \mathbb{R}^d$), a frequency $p : \Omega \to [0, 1]$ that models the prevalence of the *Wolbachia* infection trait. More specifically, in the case of cytoplasmic incompatibility caused in *Aedes* mosquitoes by the endo-symbiotic bacterium *Wolbachia*, $p$ is the proportion of mosquitoes infected by the bacterium (*e.g.* $p = 1$ means that the whole population is infected). Then, this frequency obeys a bistable reaction-diffusion equation. We aim at estimating the invasion success probability with respect to the initial data (= release profile).

In [29, 211] it was obtained an expression for the reaction term $f$ in the limit Allen-Cahn equation

$$\partial_t p - \sigma \Delta p = f(p) \tag{7.1}$$

in terms of the following biological parameters: $\sigma$ diffusivity (in square-meters per day, for example), $s_f$ (effect of *Wolbachia* on fecundity, $= 0$ if it has no effect); $s_h$ (strength of the cytoplasmic incompatibility, $= 1$ if it is perfect); $\delta$ (effect on death rate, $d_i = \delta d_s$ where $d_s$ is the regular death rate without *Wolbachia*) and $\mu$ (imperfection of vertical transmission, expected to be small). It reads as follows:

$$f(p) = \delta d_s p \frac{-s_h p^2 + \big(1 + s_h - (1 - s_f)(\frac{1-\mu}{\delta} + \mu)\big)p + (1 - s_f)\frac{1-\mu}{\delta} - 1}{s_h p^2 - (s_f + s_h)p + 1}. \tag{7.2}$$

Bistable reaction terms are such that $f < 0$ on $(0, \theta)$ and $f > 0$ on $(\theta, \theta_+)$. Usually, we consider $\theta_+ = 1$. This is the case if $\mu = 0$.

The outline of the paper is the following. In the next section, we explain how to use a threshold property for bistable reaction-diffusion equation in order to obtain explicit sufficient conditions for invasion success of a release protocol (Theorem 7.1). In a relevant stochastic framework, we show in Section 7.2.2 how these conditions provide uncertainty quantification for invasion success when release locations are random. Thanks to this, we prove in Section 7.2.3 that if the release domain is wide enough (with an explicit bound), the success probability goes to 1 as the number of releases goes to $+\infty$. Our main tool is the construction of compactly supported radially symmetric functions (in Section 7.2.4 for any dimension, and in Section 7.3 for the 1-dimensional case) such that if the initial data is above one of such functions, then invasion occurs. Section 7.3 and the following are devoted to the one dimensional case. We prove in Section 7.4.1 that the sufficient conditions for invasion are very hard to meet with a single release point (Proposition 7.5), and this leads to considering multiple release locations. For a deterministic (Section 7.4.2, Lemma 7.2) and a stochastic (Sections 7.4.3 and 7.4.4, Proposition 7.7) set of release profiles, we give analytical formulae for uncertainty quantification. Numerical simulations illustrate these results in dimension 1 in Section 7.5. We conclude in Section 7.6. Finally an appendix is devoted to the study of the minimization of the perimeter of release in one dimension.

## 7.2 Setting the problem: How to use a threshold property to design a release protocol?

### 7.2.1 The threshold phenomenon for bistable equations

In Equation (7.1), we assume that

$$\begin{cases} \exists\, \theta \in (0,1),\ f(0) = f(\theta) = f(1) = 0, \\ \\ f < 0 \text{ on } (0,\theta), \quad f > 0 \text{ on } (\theta,1), \quad \int_0^1 f(x)dx > 0. \end{cases} \tag{7.3}$$

A consequence of this hypothesis is the existence of invading traveling waves. From now on, we denote $F$ the anti-derivative of $f$ which vanishes at 0,

$$F(x) := \int_0^x f(y)\, dy. \tag{7.4}$$

Since we have assumed $F(1) > 0$, by the bistability of the function $f$, there exists a unique $\theta_c \in (0,1)$ such that

$$F(\theta_c) = \int_0^{\theta_c} f(x)dx = 0.$$

In all numerical simulations we use the following values taken from [121, 75, 181] for the *Wolbachia* and mosquito parameters:

$$\begin{aligned} d_s = 0.27\text{day}^{-1},\ s_f = 0.1,\ \mu = 0,\ s_h = 0.8, \\ \delta = 0.3/0.27 = 10/9 \text{ and } \sigma = 877\text{m}^2.\text{day}^{-1}. \end{aligned} \tag{7.5}$$

In particular we obtain the profiles for $f$ and its anti-derivative in Figure 7.1. In [121], the authors used the notations $\phi = 1 - s_f$, $\delta = d_i/d_s$, $u = 1 - s_h$ and $v = 1 - \mu$. They gave a range of values of these parameters for three *Wolbachia* strains, namely *wAlbB*, which has no impact on death ($\delta = 1$) but reduces fecundity, *wMelPop* which highly increases death rate but isn't detrimental to fecundity, and *wMel* which has a moderate impact on both. Values are given in Table 3 of the cited article (which contains also a parameter $r$, standing for differential vector competence of *Wolbachia*-infected mosquitoes for dengue, a feature we do not include in our modeling since we focus on the mosquito population dynamics), see the references therein for more details. According to the aforementioned references, the authors always assumed perfect CI and maternal transmission, that is, with our notations $s_h = 1$ and $\mu = 0$. Our notations mimic those of [29, 83], where they did not give as detailed tables for the parameters as in [121], although we refer the reader to the references they gave, which contain some quantitative estimations of these parameters. Our choices in (7.5) for $d_s, s_f$ and $\delta$ reflect the field data exposed in [75], for the (life-shortening) *wMel* strain in the context of the city of Rio de Janeiro, in Brazil.

We will always assume $\mu = 0$ (perfect vertical transmission) in the following. Complex dynamical behaviors can arise in the case when $\mu$ exceeds some threshold, as was proved in [242] for a system of two ordinary differential equations. For such values of $\mu$, in particular, population replacement may not be guaranteed by invasion success. Note however that our results apply when $\mu > 0$ is small. In this case the "invasion" state is not exactly $p = 1$, but $p = p_+(\mu) < 1$, because of the flaw in *Wolbachia* vertical (=maternal) transmission.

Moreover, following estimates from [75, 229] for *Aedes aegypti* in Rio de Janeiro (Brazil), and general literature review and discussion in Section 3 of [181] we consider that mosquitoes spread at around $\sigma = 830\text{m}^2/\text{day}$ (see the references given in [181] for more details). With these estimations of the parameters, the quantitative results we get are satisfactory because they appear to be relevant for practical purposes. For example, in order to get a significant probability of success, the release perimeter we find is around 595m wide (in one dimension). In the example from Figure 7.1, $\theta_c \simeq 0.36$.

We say that a radially symmetric function $\phi$ on $\mathbb{R}^d$ is non-increasing if $\phi(x) = g(|x|)$ for some $g$ that is non-increasing on $\mathbb{R}^+$.

The following result gives a criterion on the initial data to guarantee invasion.

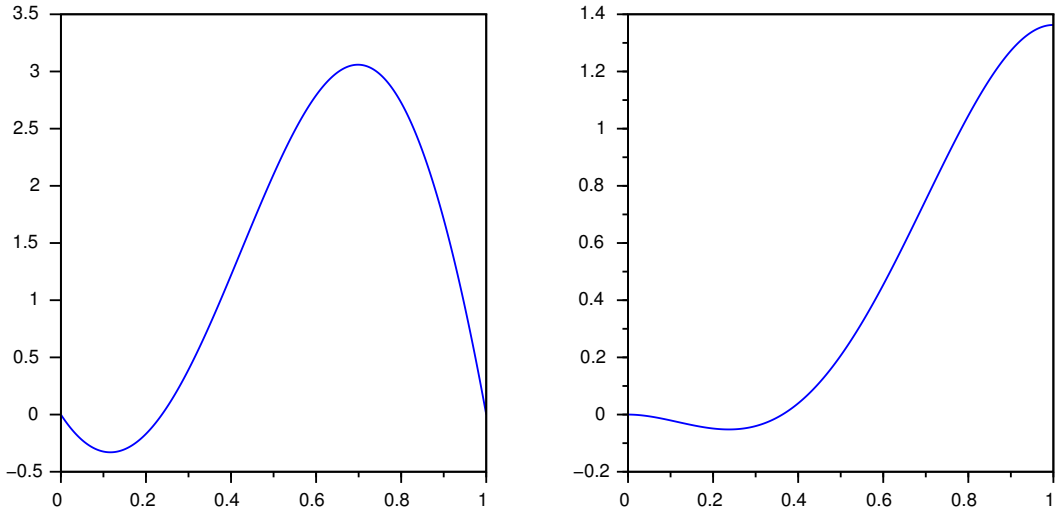



Figure 7.1: Profile of $f$ defined in (7.2) (left) and of its anti-derivative $F$ (right) with parameters given by (7.5).

**Theorem 7.1.** *Let us assume that $f$ is bistable in the sense of (7.3). Then, for all $\alpha \in (\theta_c, 1]$ there exists a compactly supported, radially symmetric non-increasing function $v_\alpha(|x|)$, with $v_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ non-increasing, $v_\alpha(0) = \alpha$ (called "$\alpha$-bubble"), such that if $p$ is a solution of*

$$\begin{aligned} &\partial_t p - \sigma \Delta p = f(p), \\ &p(t=0,x) = p^0(x) \geq v_\alpha(|x|), \end{aligned} \tag{7.6}$$

*then $p \xrightarrow[t\to\infty]{} 1$ locally uniformly. Moreover, we can take $Supp(v_\alpha) = B_{R_\alpha}$ with*

$$R_\alpha = \sqrt{\sigma} \inf_{\rho\in\Gamma} \sqrt{\frac{1-\rho^d}{(1-\rho)^2} \frac{1}{\left(\int_0^\alpha \left(1-\frac{1-\rho}{\alpha}x\right)^d f(x)dx\right)_+}}, \tag{7.7}$$

*where $\Gamma = \{\rho \in (0,1),\ \int_0^\alpha (1-\frac{1-\rho}{\alpha}x)^d f(x)dx > 0\}$.*

*In one dimension, we have the sharper estimate $Supp(v_\alpha) = [-L_\alpha, L_\alpha]$ with*

$$L_\alpha = \sqrt{\frac{\sigma}{2}} \int_0^\alpha \frac{dv}{\sqrt{F(\alpha) - F(v)}}. \tag{7.8}$$

**Remark 7.1.** *Clearly, the set $\Gamma$ is nonempty. Indeed for $\rho \sim 1$,*

$$\int_0^\alpha (1-\frac{1-\rho}{\alpha}x)^d f(x)dx > 0,$$

*since $F(\alpha) > 0$. However, it is hard to say more unless we consider a specific function $f$.*

(Sharp) threshold phenomena are well-known for bistable reaction-diffusion equations (see [70, 163, 174, 190, 243]). In Theorem 7.1, we use this property to derive the new formula (7.7), and (7.8), which are very useful to quantify invasion success uncertainty. We postpone to Section 7.2.4 the proof of this result for dimension $d \geq 1$, which is based upon an energy method developed in [174]. When $d = 1$, we give an alternative proof using sharp critical bubbles and a result of [70] in Section 7.3.1. To the best of our knowledge, we give in Section 7.3.2 the first comparison between the two approaches.

We recall the definition of a "ground state" as a positive stationary solution $v$ of (7.1), *i.e.*:

$$-\Delta v = f(v)$$

that decays to 0 at infinity. In dimension $d = 1$ (and in some special cases in higher dimensions, see [174]), such a ground state is unique up to translations. When $d = 1$ we denote $v_{\theta_c}$ the ground state which is maximal at $x = 0$. It is symmetric decreasing and $v_{\theta_c}(0) = \theta_c$, which is consistent with the notation $v_\alpha$ in Theorem 7.1. Although we won't use it in the rest of the paper, we note that with a similar argument, we have a sufficient condition for the extinction:

**Proposition 7.1.** *In dimension $d = 1$, let $p$ be a solution of equation (7.1), associated with the initial value $p_0$. If $p_0 < \theta$ or $p_0 < v_{\theta_c}(\cdot - \zeta)$ for some $\zeta \in \mathbb{R}$, then $p$ goes extinct: $p \xrightarrow[t\to\infty]{} 0$ uniformly on $\mathbb{R}$.*

### 7.2.2 The stochastic framework for release profiles

When mosquitoes are released in the field, the actual profile of *Wolbachia* infection in the days right after the release is very uncertain. In order to quantify this uncertainty, we define in this section an adequate space of release profiles. The pre-existing mosquito population is assumed to be homogeneously dense, at a level $N_0 \in \mathbb{R}_+$.

From now on, we assume that we have fixed a space unit, so that we may talk of numbers or densities of mosquitoes without any trouble.

We define a spatial process $X_\cdot(\omega) = X(\cdot, \omega) : \mathbb{R}^d \to \mathbb{R}_+$ as the introduced mosquitoes profile.

We expect that the time dynamics of the infection frequency will be given by

$$\begin{cases} \partial_t p - \sigma \Delta p = f(p), \\ \\ p(t = 0, \tau; \omega) = \dfrac{X_\tau(\omega)}{X_\tau(\omega) + N_0}. \end{cases} \tag{7.9}$$

We want to measure the probability of establishment associated with this set of initial profiles.

Making use of Theorem 7.1, we want to give a lower bound for the probability of non-extinction (which is equivalent to the probability of invasion, by the sharpness of threshold solutions, as described in [163, 174]).

An initial condition $X_\tau$ ensures non-extinction if

$$\exists \alpha \in (\theta_c, 1],\ \exists \tau_0 \in \mathbb{R},\ \forall \tau \in \mathbb{R}^d,\ \frac{X_\tau}{X_\tau + N_0} \geq v_\alpha(\tau + \tau_0), \tag{NEC}$$

where $v_\alpha$ is the "$\alpha$-bubble" used in Theorem 7.1.

**Example 1.** Now, we assume that we have a fixed number of mosquitoes to release, say $N$. When we release mosquitoes in the field (out of boxes), they will spread out to find vertebrates to feed on (if not fed in the lab prior to the release), to mate or to rest. Many environmental factors may influence their spread (see [158]). As a very rough estimate we consider that the distribution of the released mosquitoes can be described by a Gaussian. A Gaussian profile is typically the result of a diffusion process. However, we shall not use very fine properties of these profiles, and mainly focus on a "significant spread radius", so that this assumption is not too restrictive.

Due to the above simplification, the set of releases profiles ("RP") for a total of $N$ mosquitoes at $k$ locations in a domain $[-L, L]^d$ is defined as

$$RP_k^d(N) := \left\{ \tau \mapsto \frac{N}{k} \sum_{i=1}^{k} \frac{e^{-\frac{(\tau - \tau_i)^2}{2\sigma_i}}}{(2\pi\sigma_i)^{d/2}},\ \text{with } \tau_i \in [-L, L]^d,\ \sigma_i \in [\sigma_0 - \epsilon, \sigma_0 + \epsilon] \right\}, \tag{7.10}$$

where $\sigma_0$ is an estimated diffusion coefficient and $\epsilon > 0$ represents the uncertainty on this parameter ($\sigma_i$ is the "significant spread radius"). In other words, for any $i$ between 1 and $k$, the release profile is locally at the $i$-th release point a centered Gaussian with fixed amplitude $N/k$ and variance $\sigma_i$.

The basic requirement for a release profile is that $\int_{\mathbb{R}^d} X_\tau d\tau = N$. It is obviously satisfied for the elements in $RP_k^d(N)$.

We use uniform measure on $\left([-L, L]^d \times [\sigma_0 - \epsilon, \sigma_0 + \epsilon]\right)^k$ to equip $RP_k^d(N)$ with a probability measure, denoted by $\mathcal{M}$ in the following.

According to our estimate, the success probability satisfies

$$\mathbb{P}[\textit{Non-extinction after releasing } N \textit{ mosquitoes at } k \textit{ locations}\,] \geq \mathbb{P}[X_\tau(\omega) \textit{ satisfies (NEC)}], \tag{SP}$$

where $X_\tau(\omega)$ is taken in $RP_k^d(N)$ according to the uniform probability measure.

### 7.2.3 First result: relevance of under-estimating success

Though it may seem naive, our under-estimation by radii given in Theorem 7.1 is rather good, and this can be quantified in any dimension $d$. Indeed, in any dimension we can prove convergence of our under-estimation in (SP) to 1 as the number of releases goes to infinity, if we fix the number of mosquitoes per release.

More precisely, we define for a domain $\Omega \subset \mathbb{R}^d$,

$$P_k^d(N,\Omega) := \mathcal{M}\big\{(x_i)_{1\leq i\leq k}, \exists \alpha \in (\theta_c,1), \exists x_0 \in \Omega, \\ x_0 + B_{R_\alpha} \subset \Omega \text{ and } \forall x \in x_0 + B_{R_\alpha}, \frac{N}{k}\sum_{i=1}^{k} G_{\sigma,d}(x-x_i) \geq \alpha\big\}, \quad (7.11)$$

where $G_{\sigma,d}(y) = \frac{1}{(2\pi\sigma)^{d/2}} e^{-|y|^2/2\sigma}$ and $B_{R_\alpha} = B_{R_\alpha}(0)$ is the ball of radius $R_\alpha$, centered at 0. Then, the probability of success of a random (in the sense of Section 7.2.2) $k$-release of $N$ mosquitoes in the $d$-dimensional domain $\Omega$ is bigger than $P_k^d(N,\Omega)$, because of Theorem 7.1.

Fixing the number of mosquitoes per release and letting the number of releases go to $\infty$ yield:

**Proposition 7.2.** *Let $1 > \alpha > \theta_c$, $N \geq N^* := (2\pi\sigma)^{d/2}\frac{\alpha}{1-\alpha}N_0$ and $\Omega \subset \mathbb{R}^d$ be a compact set containing a ball of radius $R_\alpha$. Then,*

$$P_k^d(kN,\Omega) \xrightarrow[k\to\infty]{} 1. \quad (7.12)$$

*Proof.* There are two ingredients for the proof: First, we minimize a Gaussian at $x$ on a ball centered at $x$ by its value on the border of the ball. Second, if we pick uniformly an increasing number of balls with fixed radius and center in a compact domain, then their union covers almost-surely any given subset (this second ingredient is connected with the well-known coupon collector's problem). Namely,

$$\|y\| \leq \sqrt{2\sigma\log(2)} \implies e^{-\|y\|^2/2\sigma} \geq 1/2.$$

Now, when we pick uniformly in a compact set the centers of balls of fixed radius $\alpha$, the probability of covering a given subset $\Omega_c \subset \Omega$ increases with the number $k$ of balls. Therefore it has a limit as $k \to +\infty$. In fact, this limit is equal to 1.

One can prove this claim using the coupon collector problem (see the classical work [79] for the main results on this problem), after selecting a mesh for the compact domain $\Omega_c$. We take this mesh such that each cell has diameter less than $\sqrt{2\sigma\log(2)}/2$, and positive measure. The domain $\Omega$ is compact, hence finitely many cells is enough. Picking the center of a random ball in a given cell of the mesh has probability $> 0$, and we simply need to have picked one center in each element to be done. It remains to choose the (compact) set $\Omega_c = B_{R_\alpha} + x_0 \subset \Omega$ to conclude the proof. □

**Remark 7.2.** *We could have been a little more precise, and get an estimate for the expected value of the number $k$ of small balls required to cover the domain. According to classical results [79] on the coupon collector problem, it typically grows as $N_c\log(N_c)$, where $N_c$ is the number of cells. If the domain $\Omega$ has diameter $R$, $N_c$ is typically $(2R/\sqrt{2\sigma\log(2)})^d$, in dimension $d$.*

*Therefore we should expect $\mathbb{E}[k] \sim d\big(\frac{2R}{\sqrt{2\sigma\log(2)}}\big)^d \log\big(\frac{2R}{\sqrt{2\sigma\log(2)}}\big)$, and for a typical release area $R$ should be of the same order as $R_\alpha$.*

In fact, any $N > 0$ enjoys the same property, but then we need to assume that each cell contains a large enough number of release points.

**Corollary 7.1.** *For any $N > 0$ and $\alpha \in (\theta_c, 1)$, for $\Omega \subset \mathbb{R}^d$ a compact set containing a ball of radius $R_\alpha$, then for any compact subset $\Omega_c \subset \Omega$ containing a ball of radius $R_\alpha$ we have*

$$P_k^d(kN,\Omega_c) \xrightarrow[k\to\infty]{} 1.$$

*Proof.* Let $\iota = \lceil\!\lceil \frac{N^*}{N} \rceil\!\rceil$. With the same technique as for proving Proposition 7.2, we get a coupon collector problem where $\iota$ coupons of each kind must be collected, whence the result. □

### 7.2.4 Proof of invasiveness in Theorem 7.1 in any dimension

We consider in this section the proof of Theorem 7.1 in any dimension. The case $d = 1$ is postponed to the next section.

We use an approach based on the energy as proposed by [174]. In the present context, the energy is defined by

$$E[u] = \int_{\mathbb{R}^d} \big(\frac{\sigma}{2}|\nabla u|^2 - F(u(x))\big)dx. \tag{7.13}$$

It is straightforward to see that if $p$ is a solution to (7.6), then the energy is non-increasing along a solution, *i.e.*,

$$\frac{d}{dt}E[p] = -\int_{\mathbb{R}^d} \big(\sigma\Delta p + f(p)\big)^2\, dx \le 0.$$

Thus, $E[p](t) \le E[p^0]$ for all nonnegative $t$ and for $p$ solution to (7.6). Moreover, Theorem 2 of [174] states that if $\lim_{t\to+\infty} E[p(t,\cdot)] < 0$, then $p(t,\cdot) \to 1$ locally uniformly in $\mathbb{R}^d$ as $t \to +\infty$. Thus, since $t \mapsto E[p(t,\cdot)]$ is non-increasing, it is enough to choose $p^0$ such that $E[p^0] < 0$ to conclude the proof of Theorem 7.1.

For any $\alpha > \theta_c$, we construct $p^0(x) = v_\alpha(|x|)$ as defined in the statements of Theorem 7.1. To do so, consider the family of non-increasing radially symmetric functions, compactly supported in $B_{R_0}$ with $R_0 > 0$, indexed by a small radius $0 < r_0 < R_0$, defined by $\phi(r) = 1$ if $r \le r_0$, $\phi(r) = \frac{R_0 - r}{R_0 - r_0}$ if $r_0 < r < R_0$, and $\phi(r) \equiv 0$ if $r > R_0$.

For any $0 < r_0 < R_0$, $\phi$ is continuous and piecewise linear. We define $v_\alpha(r) = \alpha\phi(r)$, for $r \ge 0$. By the comparison principle, it suffices to find $(r_0, R_0)$ such that $E[\alpha\phi] < 0$ to ensure that $R_\alpha = R_0$ is suitable in Equation (7.7) of Theorem 7.1. To do so, we introduce

$$J_d(r_0, R_0, \alpha, \phi) := \frac{E[\alpha\phi]}{|S^{d-1}|} = \alpha^2\sigma \int_0^\infty r^{d-1}|\nabla\phi(r)|^2 dr \\ - \big(\frac{r_0^d}{d}F(\alpha) + \int_{r_0}^{R_0} r^{d-1}\int_0^{\alpha\phi(r)} f(s)dsdr\big). \tag{7.14}$$

Now, we use our specific choice of non-increasing radially symmetric function $\phi$. Introducing $\rho := r_0/R_0$, and with obvious abuses of notation, $J_d$ stands again for

$$J_d(\rho, R_0, \alpha) := R_0^d\Big(\frac{\sigma}{dR_0^2}\frac{1-\rho^d}{(1-\rho)^2} - F(\alpha)\frac{\rho^d}{d} - \frac{1-\rho}{\alpha}\int_0^\alpha \big(1 - \frac{1-\rho}{\alpha}x\big)^{d-1}F(x)dx\Big), \tag{7.15}$$

where $F$ is the antiderivative of $f$ (as introduced in (7.4)). After an integration by parts, we have

$$J_d(\rho, R_0, \alpha) = R_0^d\Big(\frac{\sigma}{dR_0^2}\frac{1-\rho^d}{(1-\rho)^2} - \int_0^\alpha \big(1 - \frac{1-\rho}{\alpha}x\big)^d f(x)dx\Big).$$

We choose $\rho \in (0,1)$ such that

$$\int_0^\alpha \big(1 - \frac{1-\rho}{\alpha}x\big)^d f(x)dx > 0 \tag{7.16}$$

Then the energy $J_d(\rho, R_0, \alpha)$ decreases to $-\infty$ with $R_0$ and is positive for $R_0 \to 0$, so the minimal scaling ensuring negative energy is obtained for some known value of $R_0 =: R_\alpha^{(d)}(\rho)$, such that $J_d(\rho, R_\alpha^{(d)}(\rho), \alpha) = 0$. Namely,

$$\big(R_\alpha^{(d)}(\rho)\big)^2 = \sigma\frac{1-\rho^d}{(1-\rho)^2}\frac{1}{\int_0^\alpha \big(1 - \frac{1-\rho}{\alpha}x\big)^d f(x)dx}, \tag{7.17}$$

which is a rational fraction in $\rho$. Thus we recover formula (7.7) in Theorem 7.1 by minimizing with respect to those $\rho$ satisfying constraint (7.16). □

We examine in particular formula (7.17) in the case $d = 1$. To do so, we introduce

$$U(\alpha) := F(\alpha) - \frac{1}{\alpha}\int_0^\alpha F(x)dx, \quad V(\alpha) := \frac{1}{\alpha}\int_0^\alpha F(x)dx. \tag{7.18}$$

Since $F(x) \leq F(\alpha)$ for $x \leq \alpha$, we know that $U$ is positive and $V$ is increasing with respect to $\alpha$ ($V'(\alpha) = \frac{1}{\alpha}U(\alpha)$). Moreover, $V(\theta_c) < 0$. We get

$$R_\alpha^{(1)}(\rho) = \frac{\alpha\sqrt{\sigma}}{\sqrt{(1-\rho)(V(\alpha)+\rho U(\alpha))}}, \tag{7.19}$$

under the constraint $V(\alpha) + \rho U(\alpha) > 0$. The optimal choice for $\rho$ is then $\rho_1^*(\alpha) := \frac{1}{2} - \frac{1}{2}\frac{V(\alpha)}{U(\alpha)}$. It satisfies $V(\alpha) + \rho_1^*(\alpha)U(\alpha) > 0$ since $U(\alpha) = F(\alpha) - V(\alpha) > 0$ and $F(\alpha) > 0$.

Finally, $\rho_1^*$ corresponds to a minimal radius

$$R_\alpha^{(1),*} := R_\alpha^{(1)}(\rho_1^*(\alpha)) = 2\sqrt{\sigma}\frac{\alpha\sqrt{U(\alpha)}}{F(\alpha)}, \tag{7.20}$$

with $U(\alpha)$ as in (7.18).

**Remark 7.3.** *We emphasize that $R_\alpha$ quantifies the minimal radius which ensures invasion from level $\alpha$, in the sense that it provides an upper bound for it. However, we were not able to perform an analytical computation of the actual optimal radius (=support size) of a critical bubble.*

**Remark 7.4.** *We note in passing that the same energy (7.13) appears for instance in the review paper [27] and in associated literature, but is used in a different spirit (stemming from statistical physics).*

Before restricting to dimension 1 in the sequel, we end the general exposition in this section with a numerical illustration. In order to help the reader getting a clearer picture of the invasion problem we investigate in the present paper, Figure 7.2 displays the time dynamics of equation (7.1) in two spatial dimensions, with three different initial conditions. In this simulation we use the function $f$ defined in (7.2) with parameter values given in (7.5). It illustrates the fact that with a fixed number of release points taken uniformly in a rectangle, invasion typically appears only if the size of the rectangle is well chosen.

If it is too small (Figure 7.2-Right) the pressure of the surrounding *Wolbachia*-free environment is too strong for the infection to propagate. If it is too large (Figure 7.2-Left), the release points are likely to be too scattered and never reach and invasion threshold. Whereas in Figure 7.2-Center, the release area and the number of releases is sufficient to generate a wide enough domain of *Wolbachia*-infected mosquitoes which spreads for larger times.

## 7.3 Critical bubbles of non-extinction in dimension 1

### 7.3.1 Construction

In this section, we consider the particular one dimensional case for which we can construct a sharp critical bubble. To do so, we consider the following differential system:

$$\sigma u_\alpha'' + f(u_\alpha) = 0 \text{ in } \mathbb{R}_+, \quad u_\alpha(0) = \alpha,\ u_\alpha'(0) = 0. \tag{7.21}$$

**Proposition 7.3.** *System (7.21) admits a unique maximal solution $u_\alpha$; it is global and can be extended by symmetry on $\mathbb{R}$ as a function of class $\mathcal{C}^2$. Moreover, if $\alpha > \theta_c$, then $L_\alpha$ defined in (7.8) is finite and $u_\alpha$ is monotonically decreasing on $\mathbb{R}_+$ and vanishes at $L_\alpha$.*

**Definition 7.1.** *For $\alpha \in (\theta_c, 1]$, we denote by an $\alpha$-bubble in one dimension the function $v_\alpha$ defined by*

$$v_\alpha(x) = u_\alpha(|x|)^+ := \max\{0, u_\alpha(|x|)\}\ .$$

From Proposition 7.3 and Definition 7.1 we have that $v_\alpha$ is compactly supported with $\text{supp}(v_\alpha) = [-L_\alpha, L_\alpha]$.

*Proof.* Local existence is granted by Cauchy-Lipschitz theorem. Then, we multiply Equation (7.21) by $u_\alpha'$,

$$\frac{\sigma}{2}\big((u_\alpha')^2\big)' + \big(F(u_\alpha)\big)' = 0,$$

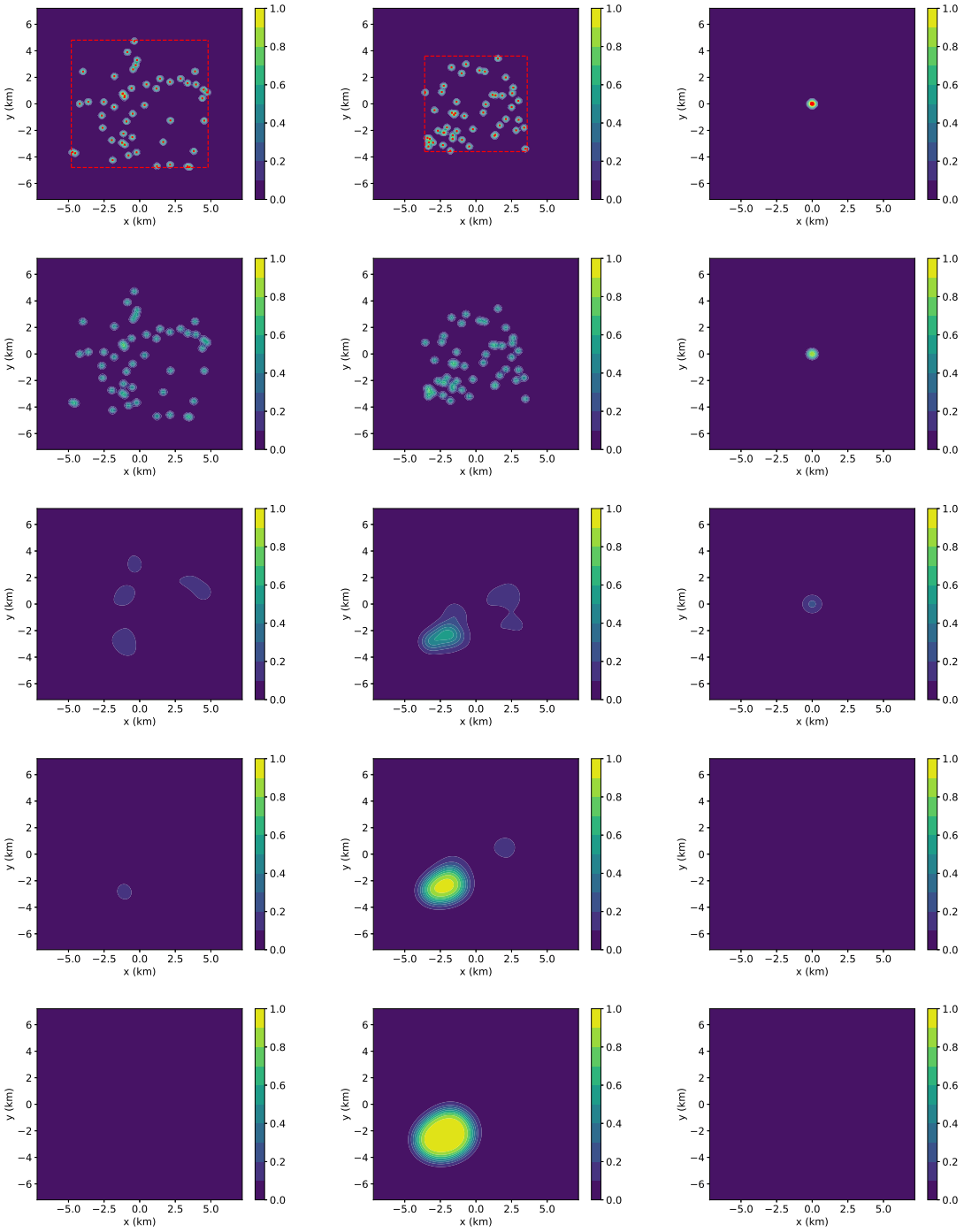



Figure 7.2: Time dynamics with three different initial releases belonging to the set $RP^2_{50}(N)$ of (7.10), with $N/(N+N_0) = 0.75$. Integration is performed on the domain $[-L, L]$ with $L = 50$km. The release box is plotted in dashed red on the first picture of each configuration. *Left*: Release box $[-2L/3, 2L/3]^2$. *Center*: Release box $[-L/2, L/2]^2$. *Right*: Release box $[-L/12.5, L/12.5]^2$. *From top to bottom*: increasing time $t \in \{0, 1, 25, 50, 75\}$, in days. The color indicates the value of $p$ (with the scale on the right).

which implies (since $u'_\alpha(0) = 0, u_\alpha(0) = \alpha$ and the domain is connected) that:

$$\frac{\sigma}{2}(u'_\alpha)^2 = F(\alpha) - F(u_\alpha).$$

Recall that $F(x) = \int_0^x f(y)dy$ is positive and increasing from $\theta_c$. Hence, for $\alpha > \theta_c$, $u_\alpha$ stays strictly below $\alpha$ except at 0; $u'_\alpha$ cannot vanish unless $u_\alpha = \alpha$. Hence, $u_\alpha$ is decreasing on $\mathbb{R}_+$.

Because $u_\alpha$ is decreasing, its derivative is negative and thus:

$$\sqrt{\sigma}\frac{du_\alpha}{dx} = -\sqrt{2(F(\alpha) - F(u_\alpha))}. \tag{7.22}$$

Then, $u_\alpha$, being monotonic, is invertible on its range. Let us define $\chi_\alpha(u_\alpha(x)) = x$, so that $u_\alpha(\chi_\alpha(\omega)) = \omega$. By the chain rule, we have

$$\frac{d\chi_\alpha}{d\omega} = -\sqrt{\frac{\sigma}{2(F(\alpha) - F(\omega))}},$$

so that,

$$\chi_\alpha(\omega) = \int_\omega^\alpha \sqrt{\frac{\sigma}{2(F(\alpha) - F(v))}}dv. \tag{7.23}$$

Thus the function $\chi_\alpha$ evaluated at $\omega$ is equal to the unique radius at which the solution of (7.21) takes the value $\omega$. It remains to prove that $L_\alpha = \chi_\alpha(0)$ is finite, *i.e.* that $v \mapsto \frac{1}{\sqrt{F(\alpha)-F(v)}}$ is integrable on $(0, \alpha)$. This function has the following equivalents at $\alpha$ and 0:

$$\frac{1}{\sqrt{F(\alpha) - F(v)}} \underset{v\to\alpha}{\sim} \frac{1}{\sqrt{f(\alpha)}}\frac{1}{\sqrt{\alpha - v}},$$

$$\frac{1}{\sqrt{F(\alpha) - F(v)}} \underset{v\to 0^+}{\sim} \begin{cases} \frac{1}{v}\sqrt{-\frac{2}{f'(0)}} & \text{if } \alpha = \theta_c, \\ \frac{1}{\sqrt{F(\alpha)}} & \text{if } \alpha > \theta_c. \end{cases}$$

Therefore $L_\alpha$ is finite if and only if $\alpha > \theta_c$. □

**Proposition 7.4.** *The limit bubble $u_{\theta_c}$ (also known as the "ground state") is exponentially decaying at infinity.*

*Proof.* The function $u_{\theta_c}$ satisfies the following equation:

$$\frac{\sigma}{2}(u'_{\theta_c})^2 = F(\theta_c) - F(u_{\theta_c}) = -F(u_{\theta_c}).$$

Hence,

$$\sqrt{\sigma}u'_{\theta_c} = -\sqrt{-2F(u_{\theta_c})} \text{ on } \mathbb{R}_+.$$

Moreover, for small $\epsilon$, $\sqrt{-2F(\epsilon)} = \epsilon\sqrt{-f'(0)} + o(\epsilon)$.

As a consequence, as $u_{\theta_c}$ gets small (at infinity), it is equivalent to the solution of

$$y' = -\sqrt{-f'(0)}y,$$

that is $x \mapsto e^{-\sqrt{-f'(0)}x}$. □

*Proof of Theorem 7.1 in dimension d=1.* Let $\alpha \in (\theta_c, 1]$, and let us assume that the initial data for system (7.1) satisfies $p(0, \cdot) \geq v_\alpha$ where $v_\alpha$ is the $\alpha$-bubble defined in Definition 7.1. From Proposition 7.3, it suffices to prove that $p(t, \cdot) \to 1$ locally uniformly on $\mathbb{R}$ as $t \to +\infty$.

We first notice that the $\alpha$-bubble $v_\alpha$ is a sub-solution for (7.1). Indeed it is the minimum between the two sub-solutions 0 and $u_\alpha$. Therefore, by the comparison principle, if $p(0, \cdot) \geq v_\alpha$, then for all $t > 0$, $p(t, \cdot) \geq v_\alpha$.

Then, the proof follows from the "sharp threshold phenomenon" for bistable equations, as exposed for example in [70, Theorem 1.3], which we recall below:

**Theorem 7.2.** *[70, Theorem 1.3] Let $\phi_\lambda$, $\lambda > 0$ be a family of $L^\infty(\mathbb{R})$ nonnegative, compactly supported initial data such that*
*(i) $\lambda \mapsto \phi_\lambda$ is continuous from $\mathbb{R}^+$ to $L^1(\mathbb{R})$;*
*(ii) if $0 < \lambda_1 < \lambda_2$ then $\phi_{\lambda_1} \leq \phi_{\lambda_2}$ and $\phi_{\lambda_1} \neq \phi_{\lambda_2}$;*
*(iii) $\lim_{\lambda \to 0} \phi_\lambda(x) = 0$ a.e. in $\mathbb{R}$.*

*Let $p_\lambda$ be the solution to (7.1) with initial data $p_\lambda(0, \cdot) = \phi_\lambda$. Then, one of the following alternative holds:*
*(a) $\lim_{t\to\infty} p_\lambda(t, x) = 0$ uniformly in $\mathbb{R}$ for every $\lambda > 0$;*
*(b) there exists $\lambda^* \geq 0$ and $x_0 \in \mathbb{R}$ such that*

$$\lim_{t\to\infty} p_\lambda(t,x) = \begin{cases} 0 & \text{uniformly in } \mathbb{R} \quad (0 \leq \lambda < \lambda^*), \\ u_{\theta_c}(x - x_0) & \text{uniformly in } \mathbb{R} \quad (\lambda = \lambda^*), \\ 1 & \text{locally uniformly in } \mathbb{R} \quad (\lambda > \lambda^*). \end{cases}$$

In our case, we define $\phi_\lambda(x) = v_\alpha(\frac{x}{\lambda})$ for $\lambda > 0$. We have $\phi_1 = v_\alpha$. Since $v_\alpha$ is a sub-solution to (7.1), the solution to this equation with initial data $\phi_1$ stays above $v_\alpha$ for all positive time. From the alternative in the above Theorem, we deduce that the solution to (7.1) with initial data $v_\alpha$ converges to 1 as time goes to $+\infty$ locally uniformly on $\mathbb{R}$. (Indeed, the ground state $u_{\theta_c}$ is bounded from above by $\theta_c < \alpha$.) By the comparison principle, we conclude that if $p(0, \cdot) \geq v_\alpha$, then $\lim_{t\to+\infty} p(t, \cdot) = 1$ locally uniformly as $t \to +\infty$. □

### 7.3.2 Comparison of the energy and critical bubble methods

Our construction of a critical $\alpha$-bubble, inspired by [29], holds in dimension 1. In this context we may compare the "minimal invasion radius" at level $\alpha$ for initial data, given by the two sufficient conditions: being above an $\alpha$-bubble (which is the maximum of two stationary solutions), or being above an initial condition with negative energy.

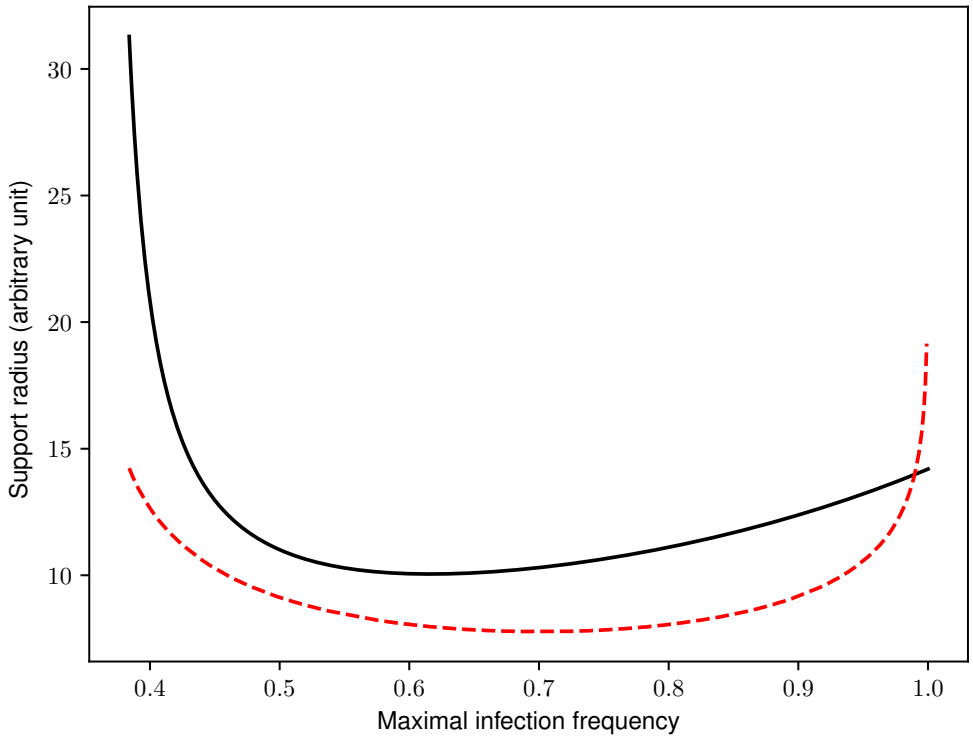



Figure 7.3: Comparison of minimal invasion radii $R_\alpha$ (obtained by energy) in dashed line and $L_\alpha$ (obtained by critical bubbles) in solid line, varying with the maximal infection frequency level $\alpha$. The scale is such that $\sigma = 1$.

We first compute the energy of the critical $\alpha$-bubble $v_\alpha$ of Definition 7.1,

$$E[v_\alpha] = \int_{\mathbb{R}} \left( \frac{\sigma}{2} |v_\alpha'|^2 - F(v_\alpha) \right) dx.$$

From Equation (7.21), we have

$$E[v_\alpha] = \int_{-L_\alpha}^{L_\alpha} \big(\sigma |v_\alpha'|^2 - F(\alpha)\big) dx = 2 \int_0^{L_\alpha} \sigma |v_\alpha'|^2 \, dx - 2L_\alpha F(\alpha).$$

Performing the change of variable $v = v_\alpha(x)$ we have

$$\int_0^{L_\alpha} |v_\alpha'|^2 dx = \int_0^{\alpha} v_\alpha'(v_\alpha^{-1}(v)) \, dv = \frac{1}{\sqrt{\sigma}} \int_0^{\alpha} \sqrt{2(F(\alpha) - F(v))} dv,$$

where we use Equation (7.22) for the last equality. Finally, using the expression of $L_\alpha$ in (7.8) we arrive at

$$E[v_\alpha] = 2\sqrt{\sigma} \int_0^{\alpha} \frac{F(\alpha) - 2F(v)}{\sqrt{2(F(\alpha) - F(v))}} \, dv.$$

To emphasize the difference between the two sufficient conditions, we observe that when $\alpha \to \theta_c$, since $F(\theta_c) = 0$, we obtain

$$E[v_{\theta_c}] = 2\sqrt{\sigma} \int_0^{\theta_c} \sqrt{-2F(v)} \, dv > 0.$$

By continuity of $\alpha \mapsto E[v_\alpha]$ we deduce

**Lemma 7.1.** *The $\alpha$-bubbles $v_\alpha$ have positive energy if $\alpha$ is close to $\theta_c$.*

**Remark 7.5.** *In particular, the energy estimate alone does not imply invasiveness of the $\alpha$-bubbles, which justifies the interest of our particular approach in one dimension. We do not claim that the "energy" or the "bubble" method is better, but we highlight the fact that they do not perfectly overlap.*

Figure 7.3 gives a numerical illustration of the fact that $\alpha$-bubbles give smaller radii at level $\alpha$, except for $\alpha \sim 1$, and at any rate provide a smaller minimal radius for invasion when the same parameters as in Figure 7.1 are used.

## 7.4 Specific study of a relevant set of release profiles

In this section we discuss a specific release protocol, with a total of $N$ mosquitoes divided equally into $k$ locations, in a space of dimension 1. It yields a release profile in the set $RP_k^d(N)$ we defined in (7.10).

### 7.4.1 Analytical study of the case of a single release

In the case of a single release ($k = 1$), we can easily describe the relationship between the mosquito diffusivity $\sigma$ and the total number of mosquitoes to release. Morally, as long as the mosquitoes diffuse they could theoretically invade (in dimension 1) by a single release, by introducing a sufficiently large amount of mosquitoes. This is the object of the next proposition:

**Proposition 7.5.** *Let $G_\sigma(\tau) := G_{\sigma,1}(\tau) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\tau^2/2\sigma}$. The following equivalent properties hold:*

*(i) There exists $\sigma_+ : \mathbb{R}_+^* \to \mathbb{R}_+^*$ such that $NG_\sigma$ satisfies (NEC) with $\tau_0 = 0$ if and only if $\sigma \in (0, \sigma_+(N)]$.*

*(ii) There exists $N_m : \mathbb{R}_+^* \to \mathbb{R}_+^*$ such that $NG_\sigma$ satisfies (NEC) with $\tau_0 = 0$ if and only if $N \geq N_m(\sigma)$.*

*Moreover, $\sigma_+$ and $N_m = \sigma_+^{-1}$ are increasing and in both cases, evolution in (7.1) with initial data $p_{\sigma,N} := \frac{NG_\sigma}{NG_\sigma + N_0}$ yields invasion by the introduced population.*

Part $(i)$ of Proposition 7.5 asserts that if we fix the total number $N$ of mosquitoes to introduce, single introduction is a failure if diffusivity is too large. Part $(ii)$ is just the converse viewpoint: if we know estimates on the diffusivity (thanks to field experiments like mark-release-recapture for example [229]), then we can define a minimal number $N_m$ of mosquitoes to introduce at a single location to succeed.

**Remark 7.6.** *If $\alpha \in (\theta_c, 1)$ makes $NG_\sigma$ satisfy (NEC) ("be above the $\alpha$-bubble"), then necessarily (evaluating at 0 to take the maximum of $G_\sigma$), $\alpha \leq \frac{N}{N+\sqrt{2\pi\sigma}N_0}$. In particular, our under-estimation of the probability is equal to 0 as soon as*

$$N < \sqrt{2\pi\sigma}N_0\frac{\theta_c}{1-\theta_c}.$$

*Equivalently, the density of mosquitoes at the center of the single release location $\frac{N}{\sqrt{2\pi\sigma}}$ should exceed $\frac{\theta_c}{1-\theta_c}N_0$ for our estimate to prove useful. (If $\theta_c = 0.8$, this is already 4 times the existing mosquito density. If $\theta_c = \frac{2}{3}$, then it is only 2 times; in the case of Figure 7.1, $\theta_c = 0.36$ and then the ratio is only 0.56).*

*Proof of Proposition 7.5.* Both the introduction profile given by the fraction $\dfrac{NG_\sigma(\tau)}{NG_\sigma(\tau)+N_0}$ and non-extinction bubbles from Theorem 7.1 built by (7.21) $(u_\alpha(\tau))_\alpha$ are symmetric, radial-decreasing functions. Instead of comparing them, we compare their reciprocals. We define $T_{\sigma,N}$ such that for all $p \in [0,\alpha]$,

$$\frac{NG_\sigma\big(T_{\sigma,N}(p)\big)}{NG_\sigma\big(T_{\sigma,N}(p)\big)+N_0} = p,$$

and $\chi_\alpha$ such that $u_\alpha(\chi_\alpha(p)) = p$. Respectively, they read

$$\begin{cases} T_{\sigma,N}(p) = \sqrt{2\sigma}\sqrt{\log\Big(\dfrac{N}{N_0\sqrt{2\pi\sigma}}\dfrac{1-p}{p}\Big)}, \\ \chi_\alpha(p) = \sqrt{\dfrac{\sigma}{2}}\displaystyle\int_p^\alpha \frac{dv}{\sqrt{F(\alpha)-F(v)}}. \end{cases} \tag{7.24}$$

By construction, the following equivalence holds

$$\forall \tau \in \mathbb{R}_+,\ \frac{NG_\sigma(\tau)}{NG_\sigma(\tau)+N_0} \geq u_\alpha(\tau) \iff \forall p \text{ s.t. } 0 \leq p \leq \alpha,\ \chi_\alpha(p) \leq T_{\sigma,N}(p).$$

Using (7.24) this rewrites as

$$\log\Big(\frac{N}{N_0\sqrt{2\pi\sigma}}\Big) \geq \Big(\int_p^\alpha \frac{dv}{2\sqrt{F(\alpha)-F(v)}}\Big)^2 - \log\Big(\frac{1-p}{p}\Big), \forall p \in [0,\alpha]. \tag{7.25}$$

From (7.25), we define

$$J_\alpha(p) := \log(p) - \log(1-p) + \Big(\int_p^\alpha \frac{dv}{2\sqrt{F(\alpha)-F(v)}}\Big)^2, \tag{7.26}$$

$$I(\sigma,N) := \log\Big(\frac{N}{\sqrt{2\pi\sigma}N_0}\Big). \tag{7.27}$$

For any given $N$, the problem we want to solve amounts at finding couples $(\alpha,\sigma)$ such that

$$\forall p \in [0,\alpha],\ J_\alpha(p) \leq I(\sigma,N). \tag{7.28}$$

We study the function $J_\alpha$. First, we note that $J_\alpha(p) \xrightarrow[p\to 0]{} -\infty$, $J_\alpha(\alpha) = \log\big(\frac{\alpha}{1-\alpha}\big)$ and it is continuous. Moreover,

$$J'_\alpha(p) = \frac{1}{p(1-p)} - \frac{1}{\sqrt{F(\alpha)-F(p)}}\int_p^\alpha \frac{dv}{2\sqrt{F(\alpha)-F(v)}},$$

and we may compute $\lim_{p\to\alpha} J'_\alpha(p) = \frac{1}{\alpha(1-\alpha)} - \frac{1}{f(\alpha)}$. Then, we can define

$$j_\alpha := \max_{p\in[0,\alpha]} J_\alpha(p), \quad j^* := \min_{\alpha\in(\theta_c,1]} j_\alpha.$$

Thus there exists $\alpha \in (\theta_c, 1]$ such that (7.25) holds if and only if $N \geq N_0\sqrt{2\pi\sigma}e^{j^*}$. This gives Proposition 7.5 $(i)$ with $\sigma_+(N) = \frac{e^{-2j^*}}{2\pi}\left(\frac{N}{N_0}\right)^2$ and Proposition 7.5 $(ii)$ with $N_m = N_0\sqrt{2\pi\sigma_+}e^{j^*}$. □

**Remark 7.7.** *With parameter values from (7.5), the expected number of mosquitoes to release is huge, since we need to have* $\frac{N_m}{N_0\sqrt{2\pi\sigma}} = e^{j^*} \simeq 7\cdot 10^{10}$ *with* $j^* \simeq 25$*, where* $\frac{N_m}{N_0\sqrt{2\pi\sigma}}$ *is the quotient between total mosquitoes to release and wild initial population in an area of typical size* $\sqrt{2\pi\sigma}$*. (This is approximately the distance diffused in* 1 *day, equal to* 72*m in this case). To obtain* $j^*$ *numerically, we used MATLAB function* `fminbnd`*. Here, the model has a clear and crucial conclusion: it is very hard to invade an area with a single, localized release. Therefore, we must model several releases (whether in time or in space). In the rest of the paper we discuss the case of releases at multiple locations at same time* $t = 0$*.*

### 7.4.2 Equally spaced releases

If we space the $k$ release points regularly in the interval $[-L_\alpha, L_\alpha]$, we want to check that (NEC) holds for

$$X_\tau = \frac{N}{k}\sum_{i=0}^{k-1} G_\sigma\big(\tau + L_\alpha(-1 + \frac{2i}{k-1})\big).$$

Within a fairly good approximation, this is the case if

$$\forall \tau \in [-L_\alpha, L_\alpha], \quad \frac{X_\tau}{X_\tau + N_0} \geq \alpha.$$

This holds in particular if

$$N \geq \widetilde{N}(k,\alpha,\sigma) = \frac{N_0\sqrt{2\pi\sigma}}{2}\frac{\alpha}{1-\alpha}ke^{\frac{L_\alpha^2}{2\sigma(k-1)^2}}.$$

If we fix $\sigma$ then we may try to find optimal $k$ and $\alpha$ in order to minimize $\widetilde{N}$. Alternatively, we can do the same, fixing $N$ or $N/k$ (number of mosquitoes per release), and find the optimal number of releases $k$.

It is straightforward, keeping in mind that $L_\alpha$ is proportional to $\sqrt{\sigma}$, that the optimal $\alpha$ here merely depends on $k$, not on $\sigma$. We may introduce

$$j^*(k) := \min_{\alpha\in(\theta_c,1)} \frac{\alpha}{1-\alpha}e^{L_\alpha^2/(2\sigma(k-1)^2)}.$$

and find the minimal (in view of our sufficient criterion) value $\widetilde{N}^*$ for $\widetilde{N}$:

**Lemma 7.2.** *For* $k$ *equally spaced releases on the line, there exists an invading release profile with* $L^1$ *norm:*

$$\widetilde{N}^*(k,\sigma) = N_0\sqrt{2\pi\sigma}\frac{k}{2}j^*(k). \tag{7.29}$$

However, we want to take into account the uncertainties and variability in the release protocol and population fixation. Namely, the release points might not be exactly equally spaced, so that introducing $\widetilde{N}^*$ mosquitoes would only give some probability of success. This is what we want to quantify now and shall be addressed in Section 7.4.3.

### 7.4.3 Multiple release locations: towards a geometric problem

When we sum several Gaussians, the profile is neither symmetric (in general), nor monotone. Therefore the previous analytical argument does not apply. However, at the cost of fixing $\sigma$ we are left with a simple geometric problem.

**First step: fixing $\sigma$ and bounding by level rather than profile.** We assume first that there is no uncertainty on $\sigma$, which is taken as $\sigma_0$ ($\epsilon = 0$ in (7.10)). As a further simplification, we shall not compare the introduction frequency profile to some $\alpha$-bubble (because it is too hard), but rather to the very simple upper bound of an $\alpha$-bubble: the characteristic function $\tau \mapsto \alpha \mathbb{1}_{-L_\alpha \leq \tau \leq L_\alpha}$.

Moreover, we assume that our $k$ release locations $(x_i)_{1\leq i\leq k}$ are within the compact set $[-L, L]$, for some $L > 0$. As above, we write

$$G_\sigma(y) := \frac{1}{\sqrt{2\pi\sigma}} e^{-y^2/2\sigma},$$

and

$$\mathcal{G} = \frac{N}{k} \sum_{i=1}^{k} G_\sigma(\cdot - x_i).$$

We define

$$P(\sigma, \frac{N}{k}, (x_i)_{1\leq i\leq k}, L_0, \alpha) := \min_{[-L_\alpha + L_0, L_\alpha + L_0]} \mathcal{G} \tag{7.30}$$

Then, the probability of success for the release of $N$ mosquitoes in a total of $k$ different sites in $[-L, L]^k$, when they all spread according to $\sigma$ diffusivity, and the initial population density was $N_0$, is given by:

$$P_k(N, L) = \mathbb{P}\Big[\exists L_0 \in \mathbb{R},\ \exists \alpha \in (\theta_c, 1),\ P(\sigma, \frac{N}{k}, (x_i)_{1\leq i\leq k}, L_0, \alpha) \geq \frac{\alpha}{1-\alpha} N_0\Big]. \tag{7.31}$$

Here, the probability $\mathbb{P}$ is taken over all the real $k$-uples $(x_l)_{1\leq l\leq k}$ such that $-L < x_1 \leq \cdots \leq x_k < L$, and $[-L, L]^k$ is equipped with the uniform measure.

**Second step: transformation into a geometric problem.** In order to get a more tractable bound, we make use of the following property:

**Proposition 7.6.** *Let $(x_i)_i \in [-L, L]^k$ with $x_1 \leq \cdots \leq x_k$. We define $\mathcal{G} = \frac{N}{k} \sum_{i=1}^{k} G_\sigma(\cdot - x_i)$.*

*If there is $\alpha \in (\theta_c, 1)$ such that*

$$\frac{N}{k} \frac{1}{\sqrt{2\pi\sigma}} \geq \frac{\alpha}{1-\alpha} N_0$$

*and $1 \leq l < m \leq k$ such that*

*(i)* $\forall l \leq j \leq m-1,\ x_{j+1} - x_j \leq 2\sqrt{2\log(2)}\sqrt{\sigma}$,

*(ii)* $x_m - x_l \geq 2L_\alpha$,

*then*

$$\frac{\mathcal{G}}{\mathcal{G} + N_0} \geq v_\alpha\big(\cdot - \frac{x_m + x_l}{2}\big)\ .$$

We notice that the constant $2\sqrt{2\log(2)} \simeq 2.35$ is optimal with this property: if two translated Gaussians centered at $x_0, x_1$ are at a distance $x_1 - x_0 = \lambda\sqrt{\sigma}$, with $\lambda > 2\sqrt{2\log(2)}$, then their sum is smaller at $\frac{x_0+x_1}{2}$ than at $x_0$.

*Proof.* This property relies on the simple computation of the sum of two $G_\sigma$s, centered at $-h$ and $h$ ($h > 0$), is greater than $G_\sigma(0)$ on $[-h, h]$ as soon as $h \leq \sqrt{2\log(2)}\sqrt{\sigma}$. Figure 7.4 illustrates this property.

Indeed, considering the sum of two Gaussian $G_\sigma$,

$$\xi(x) = \frac{1}{\sqrt{2\pi\sigma}} \Big(e^{-\frac{(x+h)^2}{2\sigma}} + e^{-\frac{(x-h)^2}{2\sigma}}\Big) = 2e^{-\frac{h^2}{2\sigma}} G_\sigma(x) \cosh(\frac{xh}{\sigma}).$$

Then, recalling that $\sigma G'_\sigma(z) = -zG_\sigma(z)$, we compute

$$\frac{1}{2} e^{\frac{h^2}{2\sigma}} \sigma \xi'(x) = -xG_\sigma(x)\cosh(\frac{xh}{\sigma}) + hG_\sigma(x)\sinh(\frac{xh}{\sigma})$$
$$\frac{1}{2} e^{\frac{h^2}{2\sigma}} \sigma^2 \xi''(x) = (h^2 + x^2 - \frac{1}{\sigma})G_\sigma(x)\cosh(\frac{xh}{\sigma}) - 2hxG_\sigma(x)\sinh(\frac{xh}{\sigma}).$$

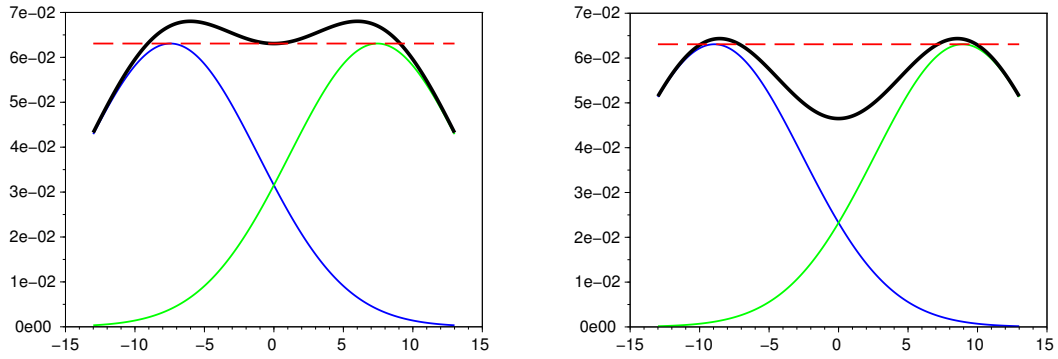



Figure 7.4: Two $G_\sigma$ profiles and their sum (in thick line). The level $G_\sigma(0)$ is the dashed line. On the left, $h = \sqrt{2\log(2)\sigma}$. On the right, $h > \sqrt{2\log(2)\sigma}$.

As a consequence, the sign of $\xi''(x)$ is that of

$$\gamma(x) := h^2 + x^2 - 2hx\tanh(\frac{xh}{\sigma}) - \frac{1}{\sigma}.$$

We notice that $\gamma(0) = h^2 - \frac{1}{\sigma}$. Hence, $\xi$ has a local maximum (resp. a local minimum) at $x = 0$ if $h < \sqrt{\sigma}$ (resp. $h > \sqrt{\sigma}$). Since $\xi(0) = 2e^{-\frac{h^2}{2\sigma}}G_\sigma(0)$, the maximal $h > 0$ that ensures $\xi(0) \geq G_\sigma(0)$ is $h = h_0 := \sqrt{2\log(2)\sigma}$.

Now, we examine the necessary condition $\xi'(x) = 0$ for a local extremum on $(-h, h)$. It implies

$$x = h\tanh(\frac{xh}{\sigma}).$$

This is true for $x = 0$ (and we have seen the condition on $h - \sqrt{\sigma}$ to have a local extremum indeed). Then, there is a solution $x_+ > 0$ if, and only if, $\frac{h^2}{\sigma} > 1$, *i.e.* $h > \sqrt{\sigma}$. In this case, $x_+$ is unique (and $x_- := -x_+$ is a solution as well).

So, for $h = h_0 > \sqrt{\sigma}$, we know that $\xi$ has a local minimum at $x = 0$, is smooth, has at most one local extremum on $(0, +\infty)$, and goes to 0 at $+\infty$. Hence, this local extremum exists and is a maximum. Therefore (and by symmetry), the minimum of $\xi$ on $(-h, h)$ is attained at $x = 0$ or $x = h$. Since $h = h_0$, $\xi(h) > \xi(0) = G_\sigma(0)$. We deduce that $\xi > G_\sigma(0)$ on $(-h, h)$.

We may use this property to prove Proposition 7.6. By condition (i) the above lower-bound holds between $x_l$ and $x_m$, and not only between two adjacent locations $x_j, x_{j+1}$. Now, the first condition implies that $G_\sigma(0) \geq \frac{\alpha}{1-\alpha}N_0$. Combining these two facts with $x_m - x_l \geq 2L_\alpha$ implies that

$$\frac{\mathcal{G}}{\mathcal{G} + N_0} \geq \alpha,$$

on $[x_l, x_m]$ which is an interval of length at least $2L_\alpha$. Precisely, for all $x \in \mathbb{R}$,

$$\frac{\mathcal{G}(x - \frac{x_m + x_l}{2})}{\mathcal{G}(x - \frac{x_m + x_l}{2}) + N_0} \geq \alpha \geq v_\alpha(x - \frac{x_m + x_l}{2}).$$

□

As a consequence, we may translate the generic inequality (SP) into:

$$P_k^1(N, (-L, L)) = P_k(N, L) \geq \mathbb{P}\Big[\exists \alpha \in (\theta_c, \frac{1}{1 + \frac{N_0}{N}k\sqrt{2\pi\sigma}}), \exists 1 \leq l < m \leq k,$$
$$x_m - x_l \geq 2L_\alpha \text{ and } \forall l \leq j \leq m-1, x_{j+1} - x_j \leq 2\sqrt{2\log(2)}\sqrt{\sigma}\Big] \quad (7.32)$$

Then, we define

$$L^* := \min_{\theta_c < \alpha \leq \frac{1}{1 + \frac{N_0}{N}k\sqrt{2\pi\sigma}}} L_\alpha,$$

and equivalently estimate (7.32) reads

$$P_k(N,L) \geq \mathbb{P}\Big[\exists 1 \leq l < m \leq k, x_m - x_l \geq 2L^* \text{ and } \max_{l \leq j \leq m-1}(x_{j+1} - x_j) \leq 2\sqrt{2\log(2)}\sqrt{\sigma}\Big]. \quad (7.33)$$

The study of the minimization of $L_\alpha$ with respect to $\alpha$ is discussed further in Appendix.

**Remark 7.8.** *Note that for this estimate, we only consider initial data that are above a characteristic function at level $\alpha$ on an interval of length $2L_\alpha$. This is far from being the optimal way to be above the $\alpha$-bubble $v_\alpha$.*

**Remark 7.9.** *It is easy to check that our estimate yields 0 (no information) as long as $k$ is too small, namely $k\sqrt{2\log(2)}\sqrt{\sigma} \leq L^*$. A necessary condition for our estimate not to yield 0 may read:*

$$k \geq \frac{1}{\sqrt{2\log(2)}} \min_{\theta_c < \alpha \leq 1} \int_0^\alpha \frac{dv}{\sqrt{2\big(F(\alpha) - F(v)\big)}}.$$

**Specific discussion for $\alpha = \theta_c$.** By Proposition 7.4, $u_{\theta_c}$ decays exponentially. As a consequence, no sum of $G_\sigma$s may be above it. This is why this profile cannot be used in our approach (because we consider that introduction profiles should be Gaussian).

### 7.4.4 Analytical computations of the probability of success: recursive formulae

In order to compute analytically the right-hand-side in (7.33), we may introduce the following notations:

- $\mathcal{T}_k(u,v)$ is the set of ordered $k$-uples between $u$ and $v$ ($u < v \in \mathbb{R}$), the measure of which is
$$\tau_k(u,v) = \frac{(v-u)_+^k}{k!}.$$
- $\mathcal{C}_k^\lambda(u,v) \subseteq \mathcal{T}_k(u,v)$ is the subset of $k$-uples such that $y_1 = u$, $y_k = v$ and for all $l \in [\![1, k-1]\!]$, $y_{l+1} - y_l \leq \lambda$. Its measure is denoted $\gamma_k^\lambda(u,v)$.
- $\mathcal{B}_k^{\lambda,R^*}(u,v) \subseteq \mathcal{T}_k(u,v)$ is the subset of $k$-uples such that $\exists 1 \leq l < m \leq k, y_m - y_l \geq R^*$ and $\max_{l \leq j \leq m-1}(y_{j+1} - y_j) \leq \lambda$. We denote $\beta_k^{L,R^*}(u,v)$ its measure.

**Remark 7.10.** *Back to problem (7.33), we recover the problem of estimating $\beta$ with the notations of Proposition 7.7 through a simple change of variables. We divide all positions $(x_1, \dots, x_k)$ by $\sqrt{2\sigma}$. Then in the right-hand side of (7.33) we replace $2L^*$ by*

$$R^* := \min_\alpha \int_0^\alpha \frac{dv}{\sqrt{F(\alpha) - F(v)}},$$

*and $2\sqrt{2\log(2)\sigma}$ by $\lambda := 2\sqrt{\log(2)}$. This was done in order to simplify computations. Moreover, it shows that the success probabilities do not depend on diffusivity. In fact, scaling in $\sigma$ as we did merely amounts at choosing a space scale such that $\sigma = 1$. Even though probabilities themselves do not make $\sigma$ appear, one must keep in mind that the corresponding release protocols (including the space between release points or the size of the release box) are proportional to $\sqrt{\sigma}$.*

We want to under-estimate the probability of success with $k$ releases in the box $[-L, L]$. In view of (SP), it amounts to computing $\frac{\beta_k^{\lambda,R^*}(-L,L)}{\tau_k(-L,L)}$. In fact, we get a general recursive formula for $\beta$ in the following proposition.

**Proposition 7.7.** *Let $k_0 := [\![\frac{R^*}{\lambda}]\!] + 1$. Then,*

$$\beta_k^{\lambda,R^*}(-L,\chi) = \sum_{i=k_0}^{k} \sum_{j=1}^{k-i+1} \int_{-L}^{\chi - R^*} \int_{u+R^*}^{\min(\chi, u+(k-1)\lambda)} \gamma_i^\lambda(u,v) \Big(\tau_{j-1}(-L, u-\lambda) - \beta_{j-1}^{\lambda,R^*}(-L, u-\lambda)\Big) \tau_{k-(i+j-1)}(v+\lambda, \chi) dv du. \quad (7.34)$$

*Proof.* The idea is simple: we count each "positive initial data", that is an ordered $k$-uple $(y_i)_i$ such that a subfamily satisfies $y_m - y_l \geq R^*$ and $y_{i+1} - y_i \leq \lambda$ between $l$ and $m$, according to its leftmost "positive sub-family", which is then taken of maximal length.

We shall use the index $i$ to denote the length of this maximal family (between $k_0$ and $k$), and $j$ its first rank ($1 \leq j \leq k - i + 1$). Then,

$$\beta_k^{\lambda,R^*}(-L,\chi) = \int_{[-L,\chi]^k} \mathbb{1}_{\{y_1 \leq y_2 \leq \cdots \leq y_k\}} \mathbb{1}_{\{(y_1,\ldots,y_k) \in \mathcal{B}_k^{\lambda,R^*}(-L,\chi)\}} dy_1 \ldots dy_k. \tag{7.35}$$

Now, we split:

$$\mathbb{1}_{\{(y_1,\ldots,y_k) \in \mathcal{B}_k^{\lambda,R^*}(-L,\chi)\}} = \sum_{i=k_0}^{k} \sum_{j=1}^{k-i+1} \mathbb{1}_{\{y_{i+j-1} - y_j \geq R^*\}} \prod_{l=j}^{j+i-2} \mathbb{1}_{\{y_{l+1} - y_l \leq \lambda\}} \mathbb{1}_{\{(y_1,\ldots,y_{j-1}) \notin \mathcal{B}_{j-1}^{\lambda,R^*}(-L,\chi)\}} \mathbb{1}_{\{y_j - y_{j-1} > \lambda\}} \mathbb{1}_{\{y_{i+j} - y_{i+j-1} > \lambda\}}. \tag{7.36}$$

This identity requires some explanations. It comes from the partition of $\mathcal{B}$ using maximal leftmost positive sub-family, as described above. Then, the term $\mathbb{1}_{\{(y_1,\ldots,y_{j-1}) \notin \mathcal{B}_{j-1}^{\lambda,R^*}(-L,\chi)\}}$ simply comes from the definition of $\mathcal{B}$. Since we consider the *leftmost* positive subfamily, no family on its left should be positive. Moreover no element on its left can be added, which justifies the $\mathbb{1}_{\{y_j - y_{j-1} > \lambda\}}$. Then, we have in addition that for $j > 1$ and $y_j \leq \chi$,

$$\mathbb{1}_{\{(y_1,\ldots,y_{j-1}) \notin \mathcal{B}_{j-1}^{\lambda,R^*}(-L,\chi)\}} \mathbb{1}_{\{y_{j-1} \leq y_j\}} \mathbb{1}_{\{y_j - y_{j-1} > \lambda\}} = \mathbb{1}_{\{(y_1,\ldots,y_{j-1}) \notin \mathcal{B}_{j-1}^{\lambda,R^*}(-L,y_j - \lambda)\}},$$

with the obvious convention that $\mathcal{B}(u,v) = \emptyset$ if $v < u$.

In addition, for $i + j - 1 < k$

$$\begin{aligned}
&\int_{[-L,\chi]^{k-(i+j-1)}} \mathbb{1}_{\{y_{i+j-1} \leq \cdots \leq y_k\}} \mathbb{1}_{\{y_{i+j} - y_{i+j-1} > \lambda\}} dy_{i+j} \ldots dy_k \\
&\qquad = \tau_{k-(i+j-1)}\big(y_{i+j-1} + \lambda, \chi\big) \\
&\qquad = \frac{\big(\chi - y_{i+j-1} - \lambda\big)_+^{k-(i+j-1)}}{\big(k - (i+j-1)\big)!}.
\end{aligned}$$

Combining these results, and using (7.35) and (7.36) yields

$$\begin{aligned}
\beta_k^{\lambda,R^*}(-L,\chi) = \sum_{i=k_0}^{k} \sum_{j=1}^{k-i+1} \int_{-L}^{\chi} \cdots \int_{x_{i+j-2}}^{\chi} \mathbb{1}_{\{y_{j+i-1} - y_j \geq R^*\}} \\
\prod_{l=j}^{j+i-2} \mathbb{1}_{\{0 \leq y_{l+1} - y_l \leq \lambda\}} \Big(\tau_{j-1}\big(-L, y_j - \lambda\big) - \beta_{j-1}^{\lambda,R^*}\big(-L, y_j - \lambda\big)\Big) \\
\tau_{k-(i+j-1)}\big(y_{i+j-1} + \lambda, \chi\big) dy_j \ldots dy_{i+j-1},
\end{aligned} \tag{7.37}$$

with conventions $\tau_0 = 1$ and $\beta_0 = 0$, regardless of their arguments.

We assume $\chi \geq -L + R^*$ (otherwise $\beta_k^{\lambda,R^*}(-L,\chi) = 0$). Using the notation $\gamma$ we introduced, Equation (7.37) is simplified again into:

$$\begin{aligned}
\beta_k^{\lambda,R^*}(-L,\chi) = \sum_{i=k_0}^{k} \sum_{j=1}^{k-i+1} \int_{-L}^{\chi-R^*} \int_{u+R^*}^{\min(\chi, u+(k-1)\lambda)} \gamma_i^{\lambda}(u,v) \\
\Big(\tau_{j-1}\big(-L, u - \lambda\big) - \beta_{j-1}^{\lambda,R^*}\big(-L, u - \lambda\big)\Big) \tau_{k-(i+j-1)}\big(v + \lambda, \chi\big) dv du,
\end{aligned}$$

where $u$ stands for $y_j$ and $v$ for $y_{i+j-1}$. This is our recursive formula (7.34). □

Now, we may give an explicit formula for $\gamma_i^{\lambda}(u,v)$. We should notice that by definition,

$$\gamma_{i+2}^{\lambda}(u,v) = \int_u^{u+\lambda} \int_{u_1}^{u_1+\lambda} \cdots \int_{u_{i-1}}^{u_{i-1}+\lambda} \mathbb{1}_{v \geq u_i \geq v-\lambda} du_i \ldots du_1,$$

that is

$$\gamma^{\lambda}_{i+2}(u,v) = \int_u^{u+\lambda} \gamma^{\lambda}_{i+1}(u_1,v)du_1. \tag{7.38}$$

Hence, we deduce the recursive formula,

**Lemma 7.3.** *For all $i, \lambda, u, v$ as above,*

$$\gamma^{\lambda}_{i+2}(u,v) = \lambda^i + \sum_{k=1}^{i+1} \frac{(-1)^k}{i!}\Big(\binom{i}{k-1}(v-u-k\lambda)^i_+ + (-1)^{i+1}\binom{i-1}{k-1}\big(k\lambda-(v-u)\big)^i_+\Big). \tag{7.39}$$

*Proof.* Obviously, $\gamma^{\lambda}_2(u,v) = \mathbb{1}_{v\geq u\geq v-\lambda}$ and we deduce from (7.38)

$$\gamma^{\lambda}_3(u,v) = \lambda + \big(v-u-2\lambda\big)_+ - \big(\lambda-(v-u)\big)_+ - \big(v-u-\lambda\big)_+$$

Then, using (7.38) again proves (7.39) by induction. □

**Remark 7.11.** *For $k < 2k_0$, formula (7.34) simplifies a lot for it is no longer recursive. It enables us to compute $\beta^{\lambda,R^*}_{k_0}(-L,L)$.*

$$\beta^{\lambda,R^*}_{k_0}(-L,L) = \int_{-L}^{L-R^*} \int_{u+R^*}^{\min(L,u+(k_0-1)\lambda)} \gamma^{\lambda}_{k_0}(u,v)dvdu. \tag{7.40}$$

*Then by (7.39) we know $\gamma^{\lambda}_{k_0}(u,v)$. With the change of variables $w = v+u$, when $L > -L+(k_0-1)\lambda$, equation (7.40) becomes*

$$\begin{aligned}\beta^{\lambda,R^*}_{k_0}(-L,L) = \int_{-L}^{L-(k_0-1)\lambda} \int_{R^*}^{(k_0-1)\lambda} \Big(\lambda^{k_0-2} + \\ \sum_{k=1}^{k_0-1} \frac{(-1)^k}{(k_0-2)!}\Big(\binom{k_0-2}{k-1}(w-k\lambda)^{k_0-2}_+ + (-1)^{k_0-1}\binom{k_0-3}{k-1}(k\lambda-w)^{k_0-2}_+\Big)\Big)dwdu \\ + \int_{L-(k_0-1)\lambda}^{L-R^*} \int_{R^*}^{L-u} \gamma^{\lambda}_{k_0}(u,u+w)dwdu.\end{aligned} \tag{7.41}$$

*Clearly, the first integral in the right-hand side of (7.41) may be written as*

$$\big(2L-(k_0-1)\lambda\big)f_1(\lambda,R^*),$$

*where $f_1$ does not depend on $L$. With the change of variables $z = L-u$, the second term in the right-hand side of (7.41) becomes*

$$\begin{aligned}f_2(\lambda,R^*) := \int_{R^*}^{(k_0-1)\lambda} \int_{R^*}^{z} \Big(\lambda^{k_0-2} + \sum_{k=1}^{k_0-1} \frac{(-1)^k}{(k_0-2)!}\Big(\binom{k_0-1}{k-1}(w-k\lambda)^{k_0-2}_+ \\ + (-1)^{k_0-1}\binom{k_0-3}{k-1}(k\lambda-w)^{k_0-2}_+\Big)\Big)dwdz.\end{aligned}$$

*In particular, it appears that it does not depend on $L$. (Recall that by definition, $k_0 = \lceil\!\lceil \frac{R^*}{\lambda} \rceil\!\rceil + 1$).*

*For $\chi \in (-L+R^*, -L+(k_0-1)\lambda)$, we can compute similarly*

$$\beta^{\lambda,R^*}_{k_0}(-L,\chi) = \int_{R^*}^{\chi-(-L)} \int_{R^*}^{z} \gamma^{\lambda}_{k_0}(0,w)dwdz,$$

*and notice that our expressions are consistent since*

$$\beta^{\lambda,R^*}_{k_0}(-L,-L+(k_0-1)\lambda) = \int_{R^*}^{-L+(k_0-1)\lambda-(-L)} \int_{R^*}^{z} \gamma^{\lambda}_{k_0}(0,w)dwdz = f_2(\lambda,R^*).$$

*All in all, $\beta_{k_0}$ is expressed as follows:*

$$\beta^{\lambda,R^*}_{k_0}(-L,\chi) = \begin{cases} 0 \text{ if } \chi+L \leq R^* \\ \displaystyle\int_{R^*}^{\chi-(-L)} \int_{R^*}^{z} \gamma^{\lambda}_{k_0}(0,w)dwdz \text{ if } \chi+L \in (R^*,(k_0-1)\lambda), \\ \big(\chi+L-(k_0-1)\lambda\big)f_1 + f_2 \text{ if } \chi+L > (k_0-1)\lambda \end{cases} \tag{7.42}$$

*(This is an affine function for $\chi + L > (k_0 - 1)\lambda$, with pent $f_1(\lambda, R^*)$).*

*Then, we obtain a bound on the probability of success with $k_0$ (the minimal number of) releases after dividing by $\tau_{k_0}(-L, L)$ :*

$$P_{k_0}(L) \geq \frac{\beta_{k_0}^{\lambda,R^*}}{\tau_{k_0}}(-L, L) = \frac{k_0!}{(2L)^{k_0}}\big((2L - (k_0 - 1)\lambda)f_1(\lambda, R^*) + f_2(\lambda, R^*)\big).$$

*In particular, we see that this underestimation of the success probability is increasing and then decreasing, and thus reaches a unique maximum at $L = \widehat{L}$.*

*We find*

$$2\widehat{L} = \lambda\big(\lceil\!\lceil \frac{R^*}{\lambda} \rceil\!\rceil + 1\big) - \frac{k_0}{k_0 - 1}\frac{f_2(\lambda, R^*)}{f_1(\lambda, R^*)}.$$

*We may note that introducing the non-negative and non-decreasing function*

$$\Gamma_k^{\lambda,R^*}(z) := \int_{R^*}^{z} \gamma_k^{\lambda}(0, w)dw$$

*we get*

$$f_1(\lambda, R^*) = \Gamma_{k_0}^{\lambda,R^*}\big((k_0 - 1)\lambda\big),$$
$$f_2(\lambda, R^*) = \int_{R^*}^{(k_0-1)\lambda} \Gamma_{k_0}^{\lambda,R^*}(z)dz.$$

*As a consequence, $f_2 \leq \big((k_0 - 1)\lambda - R^*\big)f_1$ and thus*

$$2\widehat{L} \geq \frac{k_0}{k_0 - 1}R^*.$$

## 7.5 Numerical results

Now, we present some numerical results we obtained on this set of release profiles. Numerical simulations confirm the intuition of Proposition 7.2. Our under-estimation is not very bad. Indeed, as one increases the number of release points ($k$) in a fixed perimeter, with a fixed number of mosquitoes per release, then our under-estimation of the probability of success converges to 1.

Figure 7.5 shows the probability profile as a function of the size $L$ of the release box, for 20, 40 and 80 release points. With parameter values from (7.5), $R^* = 10.981$, $\lambda = 1.665$ and thus $k_0 = 8$. The curves are obtained by a simple Monte-Carlo method. They lead to the appearance of an optimal size for the release box (6.3 in this example), that does not seem to depend on the number of release points between 20 and 80.

However, for small (relatively to $k_0$) numbers of releases, the probabilities are very small. In the case of 10 release points, the maximal probability we find is about $1.10^{-5}$.

Our numerical values are somehow consistent with field experiments: typically, the space between release points is less than $\lambda\sqrt{2\sigma}$, which is about 68m, and the optimal box size is approximately equal to $6.3 \times \sqrt{2\sigma} \simeq 257$m, with the values from (7.5).

The factor $2\sqrt{2\log(2)}$ is crucial with this respect. Losing it changes $\lambda$ from $2\sqrt{\log(2)} \simeq 1.665$ to $1/\sqrt{2} \simeq 0.707$ and makes $k_0$ ("the minimal theoretical number of releases to make our underestimation of the probability of success positive") increase from 8 to 17. We show in Figure 7.6 the probability profile for 80 releases in this case, to illustrate the loss with this "worse" geometric estimation. It culminates at around 50% only and is comparable with the green curve (for 40 release points) of Figure 7.5.

## 7.6 Conclusion and perspectives

We considered spatial aspects of a biological invasion mechanism associated to release programs and their uncertainty. We validated the framework in the one-dimensional case, and the two-dimensional case is the natural extension.

Two difficulties must be tackled in higher dimensions. First, the radial-symmetric "$\alpha$-bubbles" may still exist, but we no longer have an exact formula like (7.8) for their support. Second, the

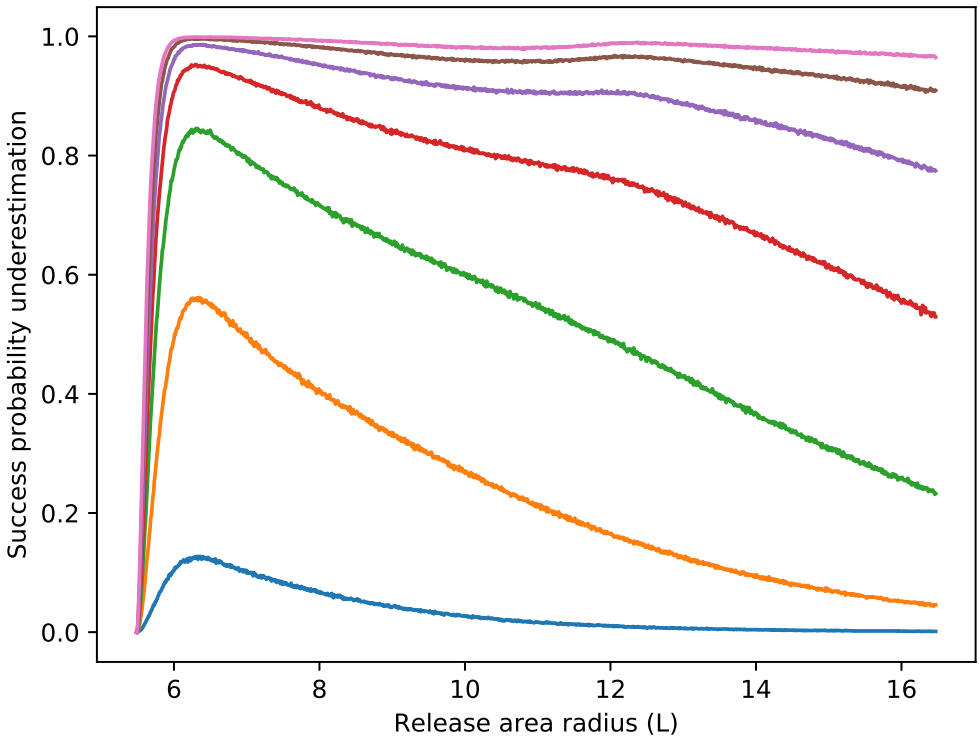



Figure 7.5: Under-estimation $\beta^{\lambda,R^*}(-L,L)$ of introduction success probability for $L$ ranging from $R^*/2 = 5.49$ to $3R^*/2 = 16.47$. The seven curves correspond to increasing number of release points. (From bottom to top: 20 to 80 release points).

geometric problem underlying our estimation gets harder, but not impossible to manage. To deal with it, we need an analogue of Proposition 7.6 in order to get a lower bound for a sum of Gaussians in two dimensions.

An interesting feature of the approach we introduced is that it can be extended to cases when neither sub-solutions nor geometric properties are available. Heuristically, we need first a criterion to tell us if a given initial data belongs to a "set of interest". Second, we need to put a probability measure on the set of "feasible initial data". Combining these, we compute the probability that the criterion is satisfied. This probability gives an insight into the role any given aspect of the release protocol plays.

We used a sufficient condition for invasion, the criterion from Theorem 7.1. However, we proved that our under-estimation of probability is rather good: in particular, it converges to 1 when the number $k$ of releases goes to $\infty$. This fact is the object of Proposition 7.2, holds true in any dimension, and is supported by numerical simulations in dimension 1.

We have always considered a homogeneous "context of introduction", so that the stochasticity would only affect the release process itself. Another natural continuation of this work, trying to go further into spatial stochasticity for release protocols, is the use of other stochastic parameters, such as the diffusion process (here it is given by a deterministic diffusivity $\sigma$), or the local carrying capacity. We let this open for further research.

Some other questions remain open. For instance: in one dimension, we considered releases in $[-L,L]$. We know that if $2L < L^*$ then our condition in the right-hand side of (7.33) is zero. On the other hand, this right-hand side goes to 0 as $L \to +\infty$. This suggests that there exists a (non-necessarily unique) size $\widehat{L}$ that maximizes this right-hand side. Back to (7.40), we obtained in Remark 7.11 a lower bound for $\widehat{L}$ in this case:

$$\widehat{L} \geq R^* \frac{1 + \lceil\!\lceil \frac{R^*}{\lambda} \rceil\!\rceil}{\lceil\!\lceil \frac{R^*}{\lambda} \rceil\!\rceil}. \tag{7.43}$$

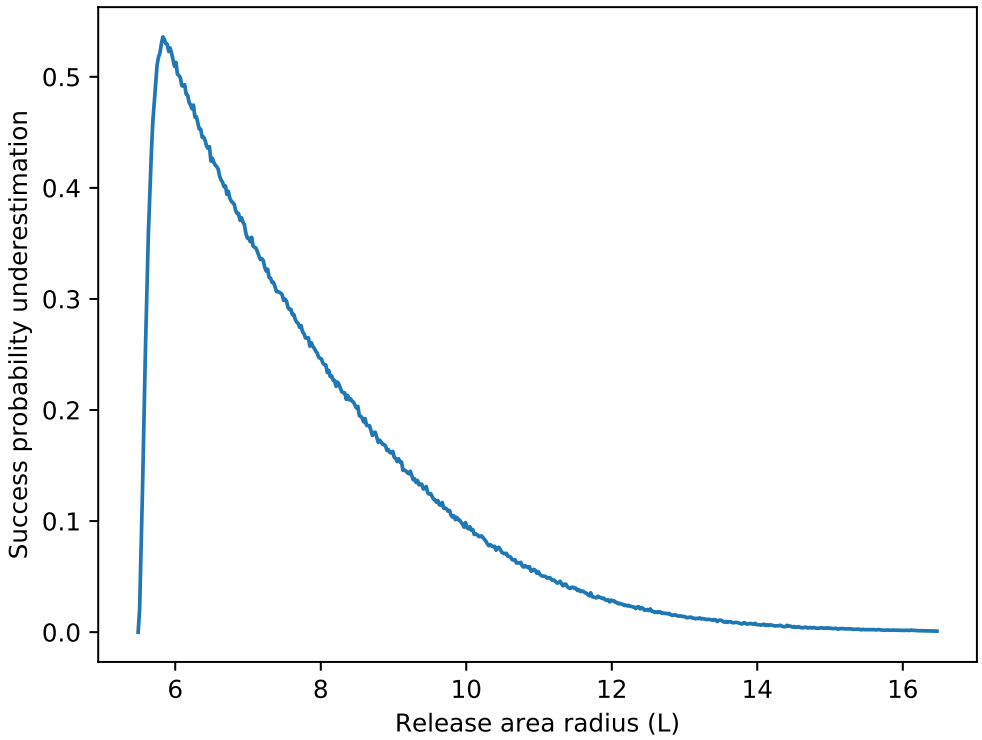



Figure 7.6: Effect of losing the constant $2\sqrt{2\log(2)}$ in Proposition 7.6: under-estimation $\beta^{\lambda,R^*}(-L,L)$ of introduction success probability for $L$ ranging from $R^*/2 = 5.49$ to $3R^*/2 = 16.47$, with 80 release points.

It is a numerical conjecture that the optimal value of $L$ is close to $\frac{1}{2}(\lambda + R^*)$ for any $k$. For this particular protocol feature (the optimal size of the release area), our approach already provides an interesting indication which - to the best of our knowledge - has not been used in previous release experiments.

As a possible follow-up to this work, one can set up several optimization problems. First, on a purely theoretical side, how to optimize the threshold functions in Theorem 7.1 with respect to a cost functional such as the $L^1$ norm (for the total number of released mosquitoes)? Then, if we fix a cost, how to maximize the under-estimated probability of success with respect to the size of the release area? Ultimately, how to optimize a release protocol (playing on the probability law of the release profiles space)?

## Acknowledgements

The authors acknowledge partial support from the Programme Convergence Sorbonne Universités / FAPERJ “Control and identification for mathematical models of Dengue epidemics” and from the Capes-Cofecub project Ma-833 15 “Modeling innovative control method for Dengue fever”. MS and NV acknowledge partial funding from the ANR blanche project Kibord: ANR-13-BS01-0004 funded by the French Ministry of Research, from the Emergence project from Mairie de Paris, Analysis and simulation of optimal shapes - application to lifesciences and from Inria, France and CAPES, Brazil (processo 99999.007551/2015-00), in the framework of the STIC AmSud project MOSTICAW. JPZ was supported by CNPq grants 302161/2003-1 and 474085/2003-1, by FAPERJ through the programs *Cientistas do Nosso Estado*, and by the Brazil-France cooperation agreement.

# Appendices

## 7.A Uniqueness of the minimal radius

In this appendix we investigate sufficient conditions for the uniqueness of a minimal radius among the $\alpha$ bubbles we constructed in Section 7.3. More precisely, we establish the number of bubbles of a given radius (which is typically 2). General results in any dimension on the exact multiplicity of solutions for such problems (semilinear elliptic Dirichlet problems) have been obtained in [182] and [183], so in essence the results below are not new and are even contained in the cited articles. However we emphasize that our proof, limited to dimension 1, uses very simple arguments and even provides an equivalent formulation of the problem in terms of a single real function $h$ built from $f$ and $F$, see formula (7.45) below.

Let $f \in \mathcal{C}^2([0,1],\mathbb{R})$ be a bistable function in the sense of (7.3) and $F(x) = \int_0^x f(y)dy$ its antiderivative as introduced in (7.4).

We make the following assumptions, with $\theta_c, \alpha_1$ defined below:

$$f'(0) < 0, \quad f'(\theta) > 0, \quad f'(1) < 0, \tag{B0}$$

$$F(1) > 0, \tag{B1}$$

$$\forall x \in [0,1], \quad \big(f'(x) + xf''(x)\big)f(x) \leq x\big(f'(x)\big)^2, \tag{B2}$$

$$\forall 1 \geq \alpha > \max(\theta_c, \alpha_1), \quad F(\alpha)\big(f(\alpha) + \alpha f'(\alpha)\big) \leq \alpha\big(f(\alpha)\big)^2. \tag{B3}$$

Under assumption (B1), there exists a unique $\theta_c \in (\theta, 1)$ such that $F(\theta_c) = 0$. We introduce

$$g(x) := xf'(x)/f(x). \tag{7.44}$$

By simple computation we have

**Lemma 7.4.** *Under assumption (B0), (B2), $g$ is decreasing on $[0,\theta)$ and on $(\theta,1]$. In addition, $g(0) = 1$, $g(\theta_-) = -\infty$, $g(\theta_+) = +\infty$ and $g(1) = -\infty$. As a consequence, there exists a unique $\alpha_1 \in (\theta,1)$ such that*

$$g(\alpha_1) = 1.$$

Now, we recall the $\alpha$-bubble radius, as introduced before, for $\alpha \in (\theta_c, 1]$:

$$L_\alpha = \sqrt{\sigma}\int_0^\alpha \frac{dv}{\sqrt{2\big(F(\alpha) - F(v)\big)}}.$$

**Proposition 7.8.** *Under conditions (B0), (B1), the bistable (in the sense of (7.3)) function $f$ is such that $L_\alpha$ reaches its minimum on $(\theta_c, 1]$ (which is well-defined) at points in $(\theta_c, 1)$.*

*If in addition (B2), (B3) hold, then there exists a unique $\alpha_0 \in (\theta_c, 1)$ such that*

$$L_{\alpha_0} = \min_\alpha L_\alpha,$$

*and for all $L > L_{\alpha_0}$, there exists unique $\alpha_\pm(L)$ with $\alpha_-(L) \in (\theta_c, \alpha_0)$ and $\alpha_+(L) \in (\alpha_0, 1)$ such that $L_{\alpha_\pm(L)} = L$.*

**Remark 7.12.** *Although assumptions (B0) and (B1) are very general, (B2) and (B3) are debatable. They yield a simple sufficient condition for uniqueness of minimum (which is the object of*

*Proposition 7.8), but are by no means necessary to get it. We expect that they can be refined and improved in order to get uniqueness for a wider class of bistable functions.*

*Using $f$ defined by (7.2) with values from (7.5), we verified numerically that (B2)-(B3) are satisfied. Indeed, using MATLAB we found that $x(f'(x))^2 - f(x)(f'(x) + xf''(x))$ and $x(f(x))^2 - F(x)(f(x) + xf'(x))$ are increasing on $[0,1]$ and $[\alpha_1, 1]$ (with $\max(\theta_c, \alpha_1) = \alpha_1$), respectively. The former is equal to 0 at 0, and the latter is approximately equal to $2 \cdot 10^{-4} > 0$ at $\alpha_1$ in this case, hence the two assumptions hold.*

*Generally, we can check that (B2)-(B3) hold for the classical bistable function $f(x) = x(1-x)(x-\theta)$ with $\theta \in (0, 1/2)$. We first compute*

$$f'(x) + xf''(x) = -9x^2 + 4(1+\theta)x - \theta.$$

*Then (B2) is equivalent to*

$$\begin{aligned}(9x^2 - 4(1+\theta)x + \theta)x(x^2 - (1+\theta)x + \theta) &\leq x(3x^2 - 2(1+\theta)x + \theta)^2 \\ \iff \\ -13(1+\theta)x^2 + 10\theta x - 5\theta(1+\theta) &\leq -12(1+\theta)x^2 + 6\theta x - 4\theta(1+\theta) \\ \iff \\ 0 &\leq (1+\theta)x^2 - 4\theta x + \theta(1+\theta).\end{aligned}$$

*The discriminant of this second-order polynomial is $-4\theta(1-\theta)^2 < 0$, so this inequality holds for any $\theta \in (0,1)$. Then a straightforard computation shows that $\alpha_1 = \frac{1+\theta}{2}$.*

*Now, we want to check (B3). To do so we compute*

$$F(x) = -\frac{1}{4}x^4 + \frac{1+\theta}{3}x^3 - \frac{\theta}{2}x^2.$$

*Then (B3) is equivalent to*

$$\begin{aligned}x^2(\frac{1}{4}x^2 - \frac{1+\theta}{3}x + \frac{\theta}{2})(4x^2 - 3(1+\theta)x + 2\theta) &\leq x^3(x^2 - (1+\theta)x + \theta)^2 \\ \iff \\ x^2(1+\theta)(2 - \frac{3}{4} - \frac{4}{3}) + \frac{\theta}{2}x + \theta(1+\theta)(2 - \frac{3}{2} - \frac{2}{3}) &\leq 0.\end{aligned}$$

*Then we recall that $2 - \frac{3}{4} - \frac{4}{3} = -\frac{1}{12} < 0$, so we just need to show that the discriminant is negative. This discriminant is equal to*

$$\frac{\theta^2}{4} - \frac{\theta(1+\theta)^2}{9} = \frac{\theta}{4}\big(\theta - \frac{4}{9}(1+\theta)^2\big) < 0.$$

*Hence simplest bistable functions of the form $f(x) = x(1-x)(x-\theta)$ satisfy our assumptions (B2) and (B3), and in particular the set of such functions is non-empty.*

*Proof.* Without loss of generality we assume $\sqrt{\sigma} = \sqrt{2}$ to get rid of the constant. From (7.8), we deduce the equivalent expression:

$$\begin{aligned}L_\alpha &= \int_0^\alpha \big(\frac{1}{\sqrt{F(\alpha) - F(v)}} - \frac{1}{\sqrt{f(\alpha)(\alpha - v)}}\big)dv + \int_0^\alpha \frac{dv}{\sqrt{f(\alpha)(\alpha - v)}} \\ &= \frac{1}{\sqrt{f(\alpha)}}\Big(\int_0^\alpha \big(\frac{\sqrt{f(\alpha)}}{\sqrt{F(\alpha) - F(v)}} - \frac{1}{\sqrt{\alpha - v}}\big)dv + 2\sqrt{\alpha}\Big)\end{aligned}$$

Hence

$$\frac{d}{d\alpha}L_\alpha = \frac{1}{\sqrt{\alpha f(\alpha)}} + \frac{1}{2\sqrt{f(\alpha)}}\int_0^\alpha \Big(\frac{1}{(\alpha - v)^{3/2}} - \big(\frac{f(\alpha)}{F(\alpha) - F(v)}\big)^{3/2}\Big)dv,$$

which is a continuous function from $(\theta_c, 1)$ to $\mathbb{R}$. It is easily seen that $\frac{d}{d\alpha}L_\alpha$ goes to $-\infty$ as $\alpha \to \theta_c^+$, and to $+\infty$ as $\alpha \to 1^-$ (recalling $f(1) = 0$). Therefore, we know that $L_\alpha$ reaches its minimum (which is well-defined) at points strictly in the interior of $(\theta_c, 1)$. This is the first point of Proposition 7.8.

Then, $\frac{d}{d\alpha}L_\alpha = 0$ if and only if

$$\frac{1}{\sqrt{\alpha}} + \frac{1}{2}\int_0^\alpha \Big(\frac{1}{(\alpha-v)^{3/2}} - \big(\frac{f(\alpha)}{F(\alpha)-F(v)}\big)^{3/2}\Big)dv = 0.$$

For $\alpha \in (\theta_c, 1)$, we introduce

$$h(\alpha) := \int_0^1 \Big(\frac{1}{(1-w)^{3/2}} - \big(\frac{\alpha f(\alpha)}{F(\alpha)-F(\alpha w)}\big)^{3/2}\Big)dw. \tag{7.45}$$

Then $\frac{d}{d\alpha}L_\alpha = 0$ if and only if $h(\alpha) = -2$. In addition, $h(\theta_c) = -\infty$ and $h(1) = +\infty$ are well-defined by continuity.

We compute

$$h'(\alpha) = -\frac{3}{2}\int_0^1 \frac{\big(\alpha f(\alpha)\big)^{1/2}}{\big(F(\alpha)-F(\alpha w)\big)^{5/2}}\Big(\big(f(\alpha)+\alpha f'(\alpha)\big)\big(F(\alpha)-F(\alpha w)\big) - \alpha f(\alpha)\big(f(\alpha)-wf(\alpha w)\big)\Big)dw,$$

and introduce

$$z(\alpha, w) := \big(f(\alpha)+\alpha f'(\alpha)\big)\big(F(\alpha)-F(\alpha w)\big) - \alpha f(\alpha)\big(f(\alpha)-wf(\alpha w)\big).$$

Now, we are going to prove that under conditions (B2), (B3), for all $\alpha \in (\theta_c, 1]$, $w \in [0,1]$,

$$z(\alpha, w) \le 0,$$

with strict inequality almost everywhere. First, we notice that $z(\alpha, 1) = 0$ and

$$z(\alpha, 0) = F(\alpha)\big(f(\alpha)+\alpha f'(\alpha)\big) - \alpha f(\alpha)^2.$$

Then we compute

$$\begin{aligned}\partial_w z &= -\alpha f(\alpha w)\big(f(\alpha)+\alpha f'(\alpha)\big) + \alpha f(\alpha)f(\alpha w) + \alpha^2 w f(\alpha)f'(\alpha w)\\ &= \alpha^2 w f(\alpha)f'(\alpha w) - \alpha^2 f(\alpha w)f'(\alpha).\end{aligned}$$

Now, denoting $g(x) = xf'(x)/f(x)$, we get

$$\partial_w z = \alpha f(\alpha w)f(\alpha)\big(g(\alpha w) - g(\alpha)\big). \tag{7.46}$$

We are going to make use of the assumptions on $f$ and equation (7.46) to prove that $z \le 0$.

Recall that there exists a unique $\alpha_1 \in (\theta, 1)$ such that $g(\alpha_1) = 1$. If $\alpha \le \alpha_1$, then for all $w \in [0, \alpha/\theta)$, $g(\alpha w) \le g(\alpha)$ while for all $w \in (\alpha/\theta, 1]$, $g(\alpha w) \ge g(\alpha)$ (these facts are stated in Lemma 7.4).

Hence $w \mapsto z(\alpha, w)$ is increasing on $[0,1]$. Since $z(\alpha, 1) = 0$, it implies that $z \le 0$.

Now, if $\alpha > \alpha_1$, there exists a unique $\beta(\alpha) \in (0, \theta)$ such that $g(\beta(\alpha)) = g(\alpha)$. In this case, if $w \in [0, \alpha/\beta(\alpha)] \cup (\theta, 1]$, $g(\alpha w) \ge g(\alpha)$. If $w \in (\alpha/\beta(\alpha), \theta)$, then $g(\alpha w) < g(\alpha)$. Hence, $\partial_w z \le 0$ on $[0, \beta(\alpha)/\alpha]$ and $\partial_w z \ge 0$ on $[\beta(\alpha)/\alpha, 1]$. It implies that $z \le 0$ if, and only if, $z(\alpha, 0) \le 0$ for all $\alpha > \alpha_1$. This is assumption (B3).

All in all, we proved that $z \le 0$ for all $\alpha, w$. Hence $h'(\alpha) > 0$, and there exists a unique $\alpha_0 \in (\theta_c, 1)$ such that $h(\alpha_0) = -2$.

We conclude that $L_\alpha$ is decreasing on $(\theta_c, \alpha_0)$ and increasing on $(\alpha_0, 1]$. Hence $\alpha_0$ is the unique minimum point of $L_\alpha$, and the uniqueness of $\alpha_\pm(L)$ follows. □

# Part III

# Temporal models

# Chapter 8

# Oscillatory regimes in a simplified model of hatching enhancement by larvae

> – Ce serait une erreur de jugement, dont je ne te crois pas capable, que de donner à un fait accidentel et, si attristant soit-il, secondaire la valeur d'un événement révélateur, lui remontra le chancelier.
> – Un fait peut être terriblement exemplaire, voire symptomatique - au reste s'agit-il d'un cas unique ? -, et je crains bien que celui-ci n'apparaisse tel [...]
>
> Jacques Abeille, *Le Veilleur du jour*.

This chapter is a joint work with Laetitia Dufour, Nicolas Vauchelet, Luis Almeida, Benoît Perthame and Daniel A.M. Villela. It originated in stimulating discussions with Claudia T. Codeço and Daniel A.M. Villela when they were visiting LJLL in June 2015, and the Hopf bifurcation part was investigated as the main topic for the master's thesis of Laetitia Dufour (March-July 2016). This work was submitted to a journal in January 2018.

**Abstract.** Understanding mosquitoes life cycle is of great interest presently because of the increasing impact of vector borne diseases in several countries. There is evidence of oscillations in mosquito populations independent of seasonality, still unexplained, based on observations both in laboratories and in nature. We propose a simple mathematical model of egg hatching enhancement by larvae which produces such oscillations that conveys a possible explanation. We propose both a theoretical analysis, based on slow-fast dynamics and Hopf bifurcation, and numerical investigations in order to shed some light on the mechanisms at work in this model.

## Introduction

Today numerous areas of the world are severely affected by mosquito-borne viral diseases, with notable examples including dengue, chikungunya and Zika (see [32]). Scientists are hard at work to find new and efficient ways to mitigate the impact of, or even eradicate these arboviral diseases, and especially target vector control.

A beneficial implementation of any of the vector control methods requires a good understanding of the local vector population's bio-ecology, and a reliable monitoring of its dynamics. To achieve better knowledge, this monitoring needs not only be demographic (using trap counts), but can also use genetic data - for the example of *Wolbachia* see [117] and [240]. However, studies in Rio de Janeiro throughout the past decade have shown that monitoring urban populations of *Aedes aegypti* is a difficult task (see [119], [75], [229]), largely because of environmental variations (spatial heterogeneity, seasonality, etc.). A first - to the best of our knowledge - systematic comparison of two complex models of *Aedes aegypti* population dynamics, relevant for a control program, was done recently in [146]. Another application of proper modeling of mosquito's life-cycle is the risk estimation for disease emergence (see [97]).

We believe that the *intrinsic* life cycle of *Aedes aegpyti* may still be improperly modeled, and effort should be put in the direction of integrating several key features in the models. Among these features, we have in mind the *transitions* between the stages (egg, larva, pupa, adult) or even within these stages (larval instars, etc.) because in theory, any of these transitions (ovipositing behavior, hatching, pupation, mating, etc.) can give rise to nonlinearity. Nonlinearities ought to be taken into account when using collected data, so that they do not blur the picture we get of the actual population's dynamics. In addition, synchronizing or de-synchronizing effects, either in time or space, are possible outputs of these nonlinearities, and can result in variations in crucial traits of the mosquito populations, such as vector capacity (see [130], [24])

We focus exclusively in this work on one single aspect of the evolution of the mosquito population, setting the hypothesis that the larval density in breeding sites directly impacts the hatching rate. Previous works on hatching and larvae dynamics include [17], [16], where stochastic models with food dynamics were used. However, to the best of our knowledge, no mathematical work has been published on the very topic of hatching enhancement through larval density since the experimental findings of [151]. Observations on this phenomenon are uneasy to obtain in the field but can be assessed in the lab (see [78]). Further research in this field could benefit from mathematical modeling tools able to take it into account and this may help monitoring the dynamics of mosquito populations.

We develop a mathematical model of the dynamics of mosquito population, with the requirements that this model be sufficiently generic to match experimental observations across various conditions and sufficiently simple so that it is possible to handle it theoretically and interpret it. Therefore we choose to develop a deterministic model based on a system of ordinary differential equations, as was done, for example, in [137]. Our simplistic model involves the positive influence of larvae on the system, acting on hatching rate. We show that this feature can explain oscillations.

We draw a general picture of the system's properties in Section 8.2, and justify rigorously the use of a two-population model as a further simplification for the identification of the qualitative properties induced by hatching feedback. Then we focus on two parameter regimes of particular interest. Firstly (Section 8.3) when the quantity of eggs is large compared to the quantity of larvae, oscillations can appear and we are faced to a slow-fast oscillatory regime giving rise to oscillation profiles comparable to those of the FitzHugh-Nagumo system (Theorem 8.1). We can compute the amplitude of the oscillations in this case, where they are typically large, and also their period. Secondly (Section 8.4), we show that our model presents a Hopf bifurcation at any positive equilibrium of the system, assuming the quantity of larvae promotes hatching. The bifurcation occurs as the feedback becomes stronger (Theorem 8.2). In this case we can compute the period of the oscillations at the bifurcation point. We provide numerical results for the system parametrized (roughly) for a tropical area such as Rio de Janeiro, showing that the range of possible oscillations is wide.

## 8.1 Models and their reduction

The life cycle of a mosquito (male and female) consists of two main stages: the aquatic stage (egg, larva, pupa), and the adult stage. We adopt a population biology point of view, which means that we describe the mosquitoes life-cycle thanks to a system of ordinary differential equations. For the purpose of studying the impact of larval density on hatching, we introduce the number densities of each population: $A(t)$ (adults), $E(t)$ (eggs), $L(t)$ (larvae) and $P(t)$ (pupae).

In a compartmental model, one can suppose the following type of dynamics

$$\begin{cases} \dfrac{d}{dt}E = \beta_E A - E\big(H(E,L) + \delta_E\big), \\ \dfrac{d}{dt}L = EH(E,L) - L\big(\phi(L) + \delta_L + \tau_L\big), \\ \dfrac{d}{dt}P = \tau_L L - \delta_P P - \tau_P P, \\ \dfrac{d}{dt}A = \tau_P P - \delta_A A. \end{cases} \tag{$S_4$}$$

We interpret the parameters as follows: $\beta_E > 0$ is the intrinsic oviposition rate; $\delta_E, \delta_L, \delta_P, \delta_A > 0$ are the death rates for eggs, larvae, pupae and adults, respectively; $\tau_L$, $\tau_P > 0$ are the transition rates from larvae to pupae and pupae to adults, respectively ; $\phi$ tunes an extra-death term due to

intra-specific competition (this term is non-linear and we assume that it depends only on the larval density); finally, $H(E,L)$ is the hatching rate, which may in general depend on larval density $L$ and on egg density $E$, neglecting a possible effect of pupae.

In order to reduce ($S_4$) to a simpler model we suppose pupa population at equilibrium. This boils down to assuming that the time dynamics for pupae is fast compared to the other compartments and thus $P = \dfrac{\tau_L}{\delta_P + \tau_P} L$.

To justify this approximation more rigorously, we assume $\tau_P, \delta_P = O(1/\epsilon)$ (quantifying the "fast dynamics" for pupae) and define $\overline{\tau}_P = \epsilon\tau_P$, $\overline{\delta}_P = \epsilon\delta_P$. We introduce $P = \epsilon M$ and then we find the following equations on $M$ and $A$ (those on $E$ and $L$ are untouched)

$$\begin{cases} \epsilon \dfrac{dM}{dt} = \tau_L L - \epsilon M(\tau_P + \delta_P), \\ \dfrac{dA}{dt} = \epsilon\tau_P M - \delta_A A. \end{cases}$$

This method follows the classical justification of Michaelis-Menten laws (see [175], [187]). We end up with

$$\begin{cases} \epsilon \dfrac{dM}{dt} = \tau_L L - M(\overline{\tau}_P + \overline{\delta}_P), \\ \dfrac{dA}{dt} = \overline{\tau}_P M - \delta_A A, \end{cases}$$

and in the limit $\epsilon \to 0$, we recover our claim under the form $M = \frac{\tau_L}{\overline{\tau}_P + \overline{\delta}_P} L$.

This simplification enables us to reduce the model to dimension 3. From now on we also assume $H(E,L) = h(L)$ and $\phi(L) = cL$ in order to obtain the simplified system

$$\begin{cases} \dfrac{d}{dt}E = \beta_E A - \delta_E E - h(L)E, \\ \dfrac{d}{dt}L = h(L)E - \delta_L L - cL^2 - \tau_L L, \\ \dfrac{d}{dt}A = \dfrac{\tau_P\tau_L}{\delta_P + \tau_P} L - \delta_A A. \end{cases} \tag{$S_3$}$$

We can proceed to a further reduction by supposing adult population at equilibrium. This boils down to assuming that the time dynamics for adult mosquitoes is fast compared to the other compartments. Exactly as above with the pupae, in the approximation when $\delta_A$ and $\beta_E$ are large (and $A$ itself is small), it makes sense to set in this system, at first order, $A = \frac{\tau_P\tau_L}{(\delta_P+\tau_P)\delta_A} L$.

Finally, system ($S_4$) reduces to the following system in dimension 2:

$$\begin{cases} \dfrac{d}{dt}E = b_E L - d_E E - h(L)E, \\ \dfrac{d}{dt}L = h(L)E - d_L L - cL^2, \end{cases} \tag{8.1}$$

where $b_E = \beta_E \frac{\tau_P\tau_L}{(\delta_P+\tau_P)\delta_A} > 0$, $d_E = \delta_E > 0$ and $d_L = \delta_L + \tau_L > 0$.

We perform this model reduction because it is sufficient to take into account the larval effect. Indeed, we show and quantify how the larval density-dependent hatching rate effectively generates oscillations, without any other source of instability (like time-delay, temperature variations or other environment-related effects). However, for future practical applications, further studies including the use of a more biologically realistic model will be mandatory.

According to experimental data (results from [78]) and mainly guided by a biological intuition we assume the hatching undergoes saturation for large values of $L$:

$$h \in \mathcal{C}^1([0,\infty)), \quad h > 0, \quad \max_L h(L) =: h_0 < +\infty. \tag{8.2}$$

For later purposes (mainly to rule out population extinction) we also assume

$$b_E > d_L + d_E, \tag{8.3}$$

$$d_E d_L < h(0)(b_E - d_L). \tag{8.4}$$

For several mosquito species, it is actually possible to identify the biological parameters $\tau_L$, $\delta_A$, $\delta_E$, $\delta_L$ and the adult density at equilibrium on the field (*i.e* $A$ such that $\frac{d}{dt}A = 0$). From the formula $L = A\frac{\delta_A}{\tau_L}$, we deduce larvae density at equilibrium on the field (this density is called $\overline{L}$ throughout this paper).

We warn the reader about what we call "equilibrium density on the field" and about parameter values. We do not claim they can precisely reproduce population variations as observed in field experiments. We simply use rough estimation of their orders of magnitude so as to prove the concept of population oscillations due to density-dependent hatching rate. See paragraph 8.2.2 for additional comments.

This warning made, from now on we consider that parameters $b_E$, $d_E$, $d_L$, adults and larvae density at equilibrium on the field are known; the competition parameter $c$ and the hatching function $h$ are unknown. The known parameters are set at a given place and temperature (see [229], [238]) and we work with a fixed temperature, so the previous biological parameters are fixed and time-independent.

Our general goal is thus to assert the possible range of remaining parameters $c$ and $h(L)$ depending on the qualitative properties of solutions.

## 8.2 Study of the reduced model

### 8.2.1 Basic properties, equilibria and their stability

With the assumption (8.2) we know that solutions remain non-negative. Furthermore, the trivial equilibrium $(0,0)$ is a steady state of (8.1) and all the other steady states $(\overline{E},\overline{L})$ are determined by a non-linear relation on $\overline{L}$

$$\begin{cases} \overline{E} = \dfrac{b_E\overline{L}}{d_E + h(\overline{L})}, \\ c\overline{L} = b_E - d_L - \dfrac{d_E b_E}{d_E + h(\overline{L})}. \end{cases} \tag{8.5}$$

We observe that solutions of (8.5) are positive if and only if $h(\overline{L}) > \frac{d_E d_L}{b_E - d_L}$. In addition:

**Lemma 8.1.** *Assume (8.2) and (8.3) hold. Then there is a constant $K > 0$ such that for all non-negative $t$, $L(t) + E(t) \leq K$. Moreover, there exists at least one positive steady state of (8.1) if and only if*

$$\min_{x\geq 0}\left(cx + \frac{d_E b_E}{d_E + h(x)}\right) \leq b_E - d_L. \tag{8.6}$$

*Furthermore, all steady states $(\overline{E},\overline{L}) \neq (0,0)$ satisfy $0 < c\overline{L} < b_E - d_L - \frac{d_E b_E}{d_E + h_0}$.*

For the first point, we do not use any property of $h$, but merely the fact that $cL^2/L \to +\infty$ as $L \to +\infty$. Note that with estimates on $h$, more restrictive properties can be obtained, in the sense that one could construct strictly smaller positively stable and attractive sets.

*Proof.* We notice that

$$\frac{d}{dt}(E+L) = b_E L - d_E E - d_L L - cL^2 \leq -d_E(E+L) + U_M,$$

where $U_M := \frac{(b_E + d_E - d_L)^2}{4c}$ is the maximum of $L \mapsto (b_E + d_E - d_L)L - cL^2$. Consequently the claim holds with $K = U_M/d_E$.

Let

$$f(x) = cx + \frac{d_E b_E}{d_E + h(x)} - (b_E - d_L).$$

Then $\overline{L}$ defines a steady state of (8.1) if and only if $f(\overline{L}) = 0$, by (8.5).

Continuity of $f$ yields the conclusion since $h_0 = \max h$. □

From now on we always assume that (8.6) holds, so that there exists at least one positive steady state of (8.1). Then we analyze the stability of those steady states.

**Lemma 8.2.** *The steady state $(0,0)$ is unstable (locally linearly) if and only if (8.4) holds.*
*A non-trivial steady state $(\overline{E},\overline{L})$ of (8.1) is unstable (locally linearly) if and only if either*

$$h'(\overline{L})\overline{E} - d_L - 2c\overline{L} - d_E - h(\overline{L}) > 0, \tag{8.7}$$

*or*

$$cd_E\overline{L} - d_E h'(\overline{L})\overline{E} + c\overline{L}h(\overline{L}) < 0 \quad \text{and} \quad h'(\overline{L})\overline{E} - d_L - 2c\overline{L} - d_E - h(\overline{L}) \leq 0. \tag{8.8}$$

*Proof.* We divide the proof into three steps.

Firstly we linearize system (8.1) around a steady state $(\overline{E},\overline{L})$. Setting $E = \overline{E} + e + \dots$ and $L = \overline{L} + \ell + \dots$, we find

$$\begin{cases} \dfrac{d}{dt}e = b_E\ell - d_E e - h(\overline{L})e - h'(\overline{L})\overline{E}\ell, \\ \dfrac{d}{dt}\ell = h(\overline{L})e + h'(\overline{L})\overline{E}\ell - d_L\ell - 2c\overline{L}\ell. \end{cases}$$

The eigenvalues $\lambda$ of the above linear system are given by the determinant

$$\begin{vmatrix} -d_E - h(\overline{L}) - \lambda & b_E - h'(\overline{L})\overline{E} \\ h(\overline{L}) & h'(\overline{L})\overline{E} - d_L - 2c\overline{L} - \lambda \end{vmatrix} = 0.$$

After straightforward computations, we obtain:

$$\lambda^2 - \lambda\Big(h'(\overline{L})\overline{E} - d_L - 2c\overline{L} - d_E - h(\overline{L})\Big) + d_E\big(d_L + 2c\overline{L} - h'(\overline{L})\overline{E}\big) + h(\overline{L})\big(d_L + 2c\overline{L} - b_E\big) = 0. \tag{8.9}$$

Secondly we look at the trivial steady-state. Taking $\overline{E} = \overline{L} = 0$ in equation (8.9), we obtain:

$$P(\lambda) := \lambda^2 + \lambda(d_L + d_E + h(0)) + d_E d_L + h(0)(d_L - b_E) = 0. \tag{8.10}$$

We are looking for the condition such that $(0,0)$ is linearly unstable (we are interested in the conditions when the mosquito population does not tend to zero in nature). In other words, we expect that the polynomial $P$ has a root with positive real part. Since the first order coefficient is positive we end up with condition (8.4) and the first point of the lemma is proved.

Finally we consider non-trivial steady states. We rewrite (8.9) as

$$\lambda^2 - \mathrm{tr}(A)\lambda + \det(A) = 0,$$

where $A$ is the Jacobian matrix of the linearized system (8.1). Using (8.5) we find

$$d_E\big(d_L + 2c\overline{L} - h'(\overline{L})\overline{E}\big) + h(\overline{L})\big(d_L + 2c\overline{L} - b_E\big) = cd_E\overline{L} - d_E h'(\overline{L})\overline{E} + c\overline{L}h(\overline{L}),$$

and thus

$$\begin{cases} \mathrm{tr}(A) = h'(\overline{L})\overline{E} - d_L - 2c\overline{L} - d_E - h(\overline{L}), \\ \det(A) = cd_E\overline{L} - d_E h'(\overline{L})\overline{E} + c\overline{L}h(\overline{L}). \end{cases} \tag{8.11}$$

The discriminant $\Delta$ of this polynomial is $\Delta = \big(\mathrm{tr}(A)\big)^2 - 4\det(A)$ and the steady state is unstable if and only if there exists a root with positive real part.

There are two cases: If $\Delta < 0$ then the real part of the roots is $\frac{\mathrm{tr}(A)}{2}$. The steady state is unstable if and only if $\mathrm{tr}(A) > 0$.

If $\Delta \geq 0$ then the bigger root is $\frac{\mathrm{tr}(A)+\sqrt{\Delta}}{2}$. Hence the steady state is unstable if and only if $\mathrm{tr}(A) > -\sqrt{\Delta}$. This is true if and only if either $\mathrm{tr}(A) > 0$ or if $\det(A) < 0$ and $\mathrm{tr}(A) \leq 0$.

□

**Remark 8.1.** *There is a link with the basic offspring number $Q_0$ (defined in [66]). This dimensionless number is the average number of offspring generated by a single fertilized mosquito: from the method in [225], we can compute $Q_0 = \sqrt{\dfrac{b_E h(0)}{d_L(d_E + h(0))}}$.*

*We remark that the first statement in Lemma 8.2 boils down to the classical property: trivial equilibrium point is unstable if and only if $Q_0 > 1$.*

**Remark 8.2.** *As in nature we can observe oscillations of eggs and larvae density [119], we pay attention in this work to oscillations around the positive steady states described in Lemma 8.2. We show in Section 8.4 that these solutions exhibit oscillations, by applying the Hopf bifurcation theorem. This behavior occurs only if the non-trivial steady state is unstable.*

For the sake of conciseness we define the following functions:

$$\begin{cases} T(k) = \dfrac{1}{\overline{L}}\Big(2k + \dfrac{k+d_E}{b_E}(k+d_E-d_L)\Big), \\ D(k) = \dfrac{1}{\overline{L}}\dfrac{k+d_E}{b_E d_E}\big(k(b_E-d_L) - d_E d_L\big). \end{cases} \tag{8.12}$$

We can rephrase Lemma 8.2 into: Let $(k,k') = (h(\overline{L}), h'(\overline{L}))$ at some equilibrium $\overline{L}$. The state $(\overline{E},\overline{L})$ is unstable if and only if either $k' > T(k)$ or $T(k) \geq k' > D(k)$. We define

$$k_\pm := \frac{d_E(b_E+2d_E+d_L) \pm \sqrt{4d_E^3(b_E-d_E-d_L) + d_E^2(b_E+2d_E+d_L)^2}}{2(b_E-d_E-d_L)}. \tag{8.13}$$

**Lemma 8.3.** *Assume (8.6) holds. If $k > k_+$ then $T(k) < D(k)$, and if $k \in (0,k_+)$ then $T(k) > D(k)$.*

*Proof.* We are looking for the $k > 0$ such that $T(k) > D(k)$, that is also written from (8.12)

$$k^2(b_E-d_E-d_L) - kd_E(b_E+2d_E+d_L) - d_E^3 < 0.$$

Recalling that $b_E > d_E + d_L$ by (8.6), the discriminant is:

$$\Delta = d_E^2(b_E+2d_E+d_L)^2 + 4d_E^3(b_E-d_E-d_L) > 0.$$

The roots are exactly $k_\pm$, so the polynomial is negative when $k \in (k_-, k_+)$.

We note that $k_- < 0$, so $T(k) < D(k)$ if and only if $k > k_+$, and $T(k) > D(k)$ if and only if $k \in (k_-,k_+)$. Since $k > 0$, this is equivalent to $k \in (0,k_+)$. □

Collecting our results on the equilibria we can state

**Proposition 8.1.** *Assume (8.6) holds, and let $(\overline{E},\overline{L})$ be a positive steady state of (8.1). Then $k_+ > \dfrac{d_E d_L}{b_E-d_L}$ and necessarily $h(\overline{L}) > \dfrac{d_E d_L}{b_E-d_L}$.*

*If $h(\overline{L}) > k_+$, then $(\overline{E},\overline{L})$ is unstable if and only if $h'(\overline{L}) > T\big(h(\overline{L})\big)$. If $\frac{d_E d_L}{b_E-d_L} < h(\overline{L}) < k_+$, then it is unstable if and only if $h'(\overline{L}) > D\big(h(\overline{L})\big)$.*

*Finally, the eigenvalues of the linearized of (8.1) at $(\overline{E},\overline{L})$ are complex conjugate and pure imaginary if and only if $h(\overline{L}) > k_+$ and $h'(\overline{L}) = T\big(h(\overline{L})\big)$.*

*Proof.* This is a direct consequence of the previous calculations, except for

$$k_+ = \frac{d_E(b_E+2d_E+d_L) + \sqrt{4d_E^3(b_E-d_E-d_L) + d_E^2(b_E+2d_E+d_L)^2}}{2(b_E-d_E-d_L)} > \frac{d_E d_L}{b_E-d_L}. \tag{8.14}$$

Inequality (8.14) is equivalent to

$$\frac{b_E-d_L}{b_E-d_L-d_E}\big(b_E+d_L+2d_E + \sqrt{(b_E+2d_E+d_L)^2 + 4d_E(b_E-d_E-d_L)}\big) > 2d_L.$$

This inequality holds because $b_E > d_L$ (thanks to (8.3)). Indeed,

$$\begin{aligned} &\frac{b_E-d_L}{b_E-d_L-d_E}\big(b_E+d_L+2d_E + \sqrt{(b_E+2d_E+d_L)^2 + 4d_E(b_E-d_E-d_L)}\big) \\ &\qquad > \big(b_E+d_L+2d_E + \sqrt{(b_E+2d_E+d_L)^2 + 4d_E(b_E-d_E-d_L)}\big) > b_E+d_L > 2d_L. \end{aligned}$$

Then, setting $k = h(\overline{L})$, $k' = h'(\overline{L})$ and using the notations (8.11), the eigenvalues of the linearized operator are roots of the polynomial

$$P(\lambda) = \lambda^2 - \lambda \mathrm{tr}(A) + \det(A).$$

Hence the roots are pure imaginary if and only if $\mathrm{tr}(A) = 0$ and $\det(A) > 0$. From the definition of $T, D$ in (8.12), $\mathrm{tr}(A) = 0$ if and only if $k' = T(k)$. As $\det(A) > 0$ if and only if $k' < D(k)$, by Lemma 8.3 this holds whenever $k > k_+$.

□

### 8.2.2 Discussion on the nonlinearities and the equilibrium values

We discuss in this paragraph the nonlinearities of system (8.1), and the role they play.

First we justify the use of a competition term. Solutions of (8.1) are bounded (Lemma 8.1), but this holds only thanks to the nonlinear competition term $-cL^2$ in the equation describing the larvae dynamics. More generally, any competition term $\phi(L)$, as in Section 8.1 such that $\phi(L) \to +\infty$ as $L \to +\infty$ yields the same result. However, in the absence of such a competition, *a priori* bound on the solutions cannot be obtained, and no phenomenon keeps the population finite. For *Aedes* mosquitoes, the amount of available food in the breeding sites is an actual resource limitation that can trigger massive death of larvae if the amount of food per larva drops down too low (see [16]). Therefore, we choose the simplest (*i.e.* quadratic) competition term to represent this competition for resources, and this ensures mathematically that solutions remain bounded.

Still, the competition parameter $c$ is extremely hard to assess from experimental data, and the values we use in this work should be handled with care. Usually, we fix a value for a positive equilibrium $\overline{L}$ (which corresponds to choosing a type of breeding site). Then, to each value $k = h(\overline{L})$ corresponds a non-necessarily unique $c(k)$ that makes $\overline{L}$ an equilibrium of (8.1). We treat $k$ as a free parameter in this study. It has been observed that the hatching rate indeed is extremely dispersed (see for instance the experimental results of [151]), depending not only on the mosquito population and the environmental conditions but also on the egg batches themselves. In future works expanding on the simplest oscillatory behavior we describe here, this variability in the actual value of $k$ should be taken into account if the model outputs are to be linked with experimental data.

Second, we discuss the hatching rate function $h$, which is crucial to our study. From now on, we require $h$ to be increasing. Indeed, Proposition 8.1 shows that a steady state is always stable if $h$ is decreasing. Hence only an increasing $h$ can produce stable oscillations. This mathematical assumption is supported by a simple biological hypothesis: larvae promote hatching.

An interesting feature of this intuition is that it can be subsequently extended to higher-dimensional systems such as ($S_4$). In other words, it is not an artifact produced by considering only a 2-dimensional system but a robust qualitative property for these systems.

Indeed, for ($S_4$) the Jacobian matrix at any point $X = (E, L, P, A)$ reads

$$J(X) = \begin{pmatrix} -\delta_E - h(L) & -Eh'(L) & 0 & \beta_E \\ h(L) & h'(L)E - \delta_L - \tau_L - 2cL & 0 & 0 \\ 0 & \tau_L & -\delta_P - \tau_P & 0 \\ 0 & 0 & \tau_P & -\delta_A \end{pmatrix},$$

hence if $h'(L) < 0$ then $J(X)$ is a Metzler matrix (it has positive extra-diagonal coefficients): the system is cooperative in this case. Its characteristic polynomial may be written

$$P(\lambda) = (\lambda + A_1)(\lambda + A_2)(\lambda^2 + A_3\lambda + A_4) - C,$$

where $A_i, C > 0$. Being a Metzler matrix, $J$ has a real dominant eigenvalue. This matrix is stable if and only if this eigenvalue is negative; in other words, if and only if $P(0) > 0$ (since $P$ is increasing on $(0, +\infty)$). This condition reads

$$\delta_A(\delta_P + \tau_P)\big((\delta_E + h(L))(-h'(L)E + \tau_L + \delta_L + 2cL) + Eh(L)h'(L)\big) > \beta_E\tau_L\tau_P h(L).$$

At equilibrium,

$$\delta_L + \tau_L + cL = \frac{h(L)}{h(L) + \delta_E}\frac{\beta_E\tau_L\tau_P}{\delta_A(\delta_P + \tau_P)},$$

therefore $P(0) > 0$ and thus any equilibrium where $h' < 0$ must be (locally) stable, in system ($S_4$) as well as in system (8.1). Adding "neutral" compartments keeps this property true and we can be confident in concluding that only a positive effect of larvae on hatching rate can destabilize the equilibrium and lead to (local) oscillations.

Some preliminary experiments ran by one of the authors seem to indicate that the larval impact on hatching may depend on larval development stage. Taking this into account would require to make the model more complex. For instance, to model hatching impact discrepancies between first instar (positive) and last instar larvae (negative) we could add at least one compartment in (8.1). However, we focus here on the simplest oscillations-producing mechanism. The hatching function

being increasing and bounded, it is reasonable to assume that $h$ is S-shaped and smooth, which is what we use in the rest of the paper.

Third, having discussed the two nonlinearities in (8.1), we are left with an important question about steady states: how to ensure that $\overline{L}$ is actually unique? The second equation in (8.5) is also written

$$h(\overline{L}) = d_E \frac{d_L + c\overline{L}}{b_E - d_L - c\overline{L}}. \tag{8.15}$$

The number of positive steady states depends strongly on function $h$. Being a S-shaped function does not guarantee uniqueness. Therefore, it should be checked case by case except for some simple function families. We illustrate this fact in next subsection with Hill functions. Still, we notice that $\kappa : L \mapsto d_E \dfrac{d_L + cL}{b_E - d_L - cL}$ is convex on $(0, (b_E - d_L)/c)$ and goes to $+\infty$ at $(b_E - d_L)/c$. So for instance uniqueness is guaranteed if (8.4) holds and either, for all $L \in (0, (b_E - d_L)/c)$, $h''(L) < 0$ or

$$h'(L) < \kappa'(L) = \frac{d_E c b_E}{(b_E - d_L - cL)^2}.$$

### 8.2.3 Observations on a class of hatching functions

Among the many possible choices for a S-shaped hatching function $h(\,\cdot\,)$, we numerically and theoretically explore the typical family of Hill functions. We assume the following form

$$h(L) = h_m + a\frac{L^p}{\lambda^p + L^p}, \quad h_m > \frac{d_E d_L}{b_E - d_L}, \tag{8.16}$$

with the parameters $a, \lambda, p > 0$.

Steady states $(\overline{E}, \overline{L})$ of (8.1) are such that $\overline{L}$ is a solution of $Q(L) = 0$, where

$$Q(L) = -cL^{p+1}\big(h_m + a + d_E\big) + L^p\big((h_m + a)(b_E - d_L) - d_E d_L\big) \\ - c\lambda^p L(h_m + d_E) + \lambda^p\big(h_m(b_E - d_L) - d_E d_L\big).$$

The following lemma is a straightforward consequence of this computation

**Lemma 8.4.** *When $p = 1$ and $h$ is of type (8.16), there is a unique steady state of (8.1).*

When $p = 1$ and $h$ is of type (8.16), then $h'' < 0$, a property that is lost when $p > 1$. Therefore, to simplify the choice of the parameters, now we assume

$$d_E = 0. \tag{8.17}$$

Then, condition (8.6) is fulfilled, the steady state of (8.1) is unique and is given by

$$\overline{L} = \frac{b_E - d_L}{c}, \quad \overline{E} = \frac{b_E \overline{L}}{h(\overline{L})}.$$

**Proposition 8.2.** *Let $h$ be of type (8.16) and assume condition (8.17) holds. Then (8.1) has a unique positive steady state $(\overline{E}, \overline{L})$ and its linearization has eigenvalues with negative real parts if and only if*

$$k > \frac{a}{1 + \alpha^p} > \frac{\alpha^p + 1}{p\alpha^p} k\big(2 + \frac{k - d_L}{b_E}\big), \quad \alpha = \frac{\lambda}{\overline{L}}, \quad k = h(\overline{L}). \tag{8.18}$$

*Proof.* Necessarily $k > k_+ = 0$ (where $k_+$ is defined in (8.13)). Hence the eigenvalues of the linearized system at $(\overline{E}, \overline{L})$ are in $\mathbb{C}\backslash\mathbb{R}$ and the condition for instability of the steady state from Proposition 8.1 simply reads $h'(\overline{L}) > T(h(\overline{L}))$, where $T$ is defined in (8.12). Then we compute

$$h(\overline{L}) = h_m + \frac{a}{1 + \alpha^p}, \quad h'(\overline{L}) = \frac{ap\alpha^p}{\overline{L}\big(\alpha^p + 1\big)^2}.$$

The right-hand side inequality in (8.18) comes from $h'(\overline{L}) > T(k)$ and the left-hand side from $h_m > 0$. □

If all parameters but $\alpha$ and $a$ are fixed, then condition (8.18) can be fulfilled if and only if

$$k < (p-2)b_E + d_L. \tag{8.19}$$

Indeed, we need to find $\alpha > 0$ such that $2 + \frac{k-d_L}{b_E} < p\frac{\alpha^p}{1+\alpha^p}$. Note that in particular, this is impossible when $p \leq 1$ (since $c\overline{L} = b_E - d_L > 0$ by hypothesis).

We refer to Appendix 8.A for numerical results showing consistent oscillations under condition (8.19), for 2- and 3-dimensional systems (8.1) and ($S_3$).

## 8.3 The slow-fast oscillatory regime

In order to understand periodic solutions to (8.1), we examine a possible regime with a small parameter and then prove the oscillation result (Theorem 8.1). We have in mind here the analysis of the FitzHugh-Nagumo system. Numerical illustration, amplitude and period computation in some particular cases can be found in Appendix 8.B.

### 8.3.1 Parameter regime and main result

Here, we assume that the egg stock is large, and its dynamics slow compared with the larvae stock. This identifies a small parameter leading to a slow-fast system.

More precisely, let $\epsilon > 0$, $\eta : \mathbb{R}_+ \to \mathbb{R}_+$, and assume at first that all parameters (except for $h$) may depend on $\epsilon$. We transform the variables $(E, L)$ from (8.1) into $v_\epsilon := \epsilon E$ and $u_\epsilon := \frac{1}{\eta(\epsilon)}L$. These new variables satisfy

$$\begin{cases} \dot{v}_\epsilon = \epsilon\eta(\epsilon)b_E u_\epsilon - \big(d_E + h(\eta(\epsilon)u_\epsilon)\big)v_\epsilon =: f_\epsilon(u_\epsilon, v_\epsilon), \\ \epsilon\dot{u}_\epsilon = \frac{1}{\eta(\epsilon)}h(\eta(\epsilon)u_\epsilon)v_\epsilon - d_L\epsilon u_\epsilon - c\eta(\epsilon)\epsilon u_\epsilon^2 =: g_\epsilon(u_\epsilon, v_\epsilon). \end{cases} \tag{8.20}$$

We assume that parameters scale in such a way that the following limits exist, as $\epsilon \to 0$:

$$\begin{cases} f_\epsilon \xrightarrow{L^\infty} f, \quad g_\epsilon \xrightarrow{L^\infty} g, \\ u_\epsilon(t=0) = u_\epsilon^0 \to u_0, \quad v_\epsilon(t=0) = v_\epsilon^0 \to v_0. \end{cases} \tag{8.21}$$

In addition, we assume that the zero set of $g$ is "non-degenerate" in the sense:

$$\forall v \geq 0, \quad \big\{\sigma \geq 0,\, g(\sigma, v) = 0\big\} \text{ does not contain any open interval.} \tag{8.22}$$

We give below a simple proof of the following fact, in the spirit of Tikhonov's theorem on dynamical systems [90].

**Theorem 8.1.** *Consider system (8.20) with $d_E, d_L, \overline{L}$ and $h$ fixed, $b_E(\epsilon) = \frac{h(\overline{L})+d_E}{\epsilon\overline{L}}$, $\eta(\epsilon) = \frac{\overline{L}^2}{h(\overline{L})-\epsilon d_L\overline{L}}$, for $\epsilon$ small enough, and $c_\epsilon = \frac{1}{\epsilon\eta(\epsilon)}$. Let $\overline{E}(\epsilon) := 1/\epsilon$. Then $(\epsilon\overline{E}(\epsilon), \frac{1}{\eta(\epsilon)}\overline{L}) = (1, \frac{h(\overline{L})-\epsilon d_L\overline{L}}{\overline{L}})$ is a steady state of (8.20) for all $\epsilon > 0$ and (8.21) holds.*

*In addition, solutions of system (8.20) along with any bounded initial data admits a limit as $\epsilon \to 0$: there exists $u, v \in L^1 \cap L^\infty(0,T)$ for all $T > 0$ such that $v_\epsilon \to v$ uniformly and $u_\epsilon \to u$ in $L^p(0,T)$ for all $p < \infty$.*

*Moreover, if initial data $u_\epsilon^0, v_\epsilon^0$ are such that $\big(sgn(g_\epsilon), sgn(f_\epsilon)\big)(u_\epsilon^0, v_\epsilon^0)$ is constant for $\epsilon$ small enough, then $(u, v)$ is periodic, $g(u(t), v(t)) = 0$ for almost every $t > 0$ and the trajectory is uniquely defined from $f$ and $g$ with $\frac{dv}{dt} = f(u, v)$.*

Figure 8.1 illustrates the slow-fast dynamics. Before proving Theorem 8.1, we justify the particular scaling choices in its statement. Non-trivial equilibrium $(\overline{E}, \overline{L})$ of (8.1) are given by (8.5)

$$d_L + c\overline{L} = \frac{h(\overline{L})b_E}{h(\overline{L}) + d_E}, \quad \overline{E} = \frac{b_E\overline{L}}{h(\overline{L}) + d_E}.$$

Thus in all generality (allowing all parameters to depend on $\epsilon$), the scalings fit for our purpose (*i.e.* with $\overline{E}(\epsilon) = 1/\epsilon$) are exactly those for which $\epsilon = \frac{h(\overline{L}(\epsilon))+d_E(\epsilon)}{b_E(\epsilon)\overline{L}(\epsilon)}$ and there exists $\eta(\epsilon) = O(1)$ such that

$$d_L(\epsilon)\big(h(\overline{L}(\epsilon)) + d_E(\epsilon)\big) + \frac{b_E(\epsilon)\overline{L}^2(\epsilon)}{\eta(\epsilon)} = h(\overline{L}(\epsilon))b_E(\epsilon).$$

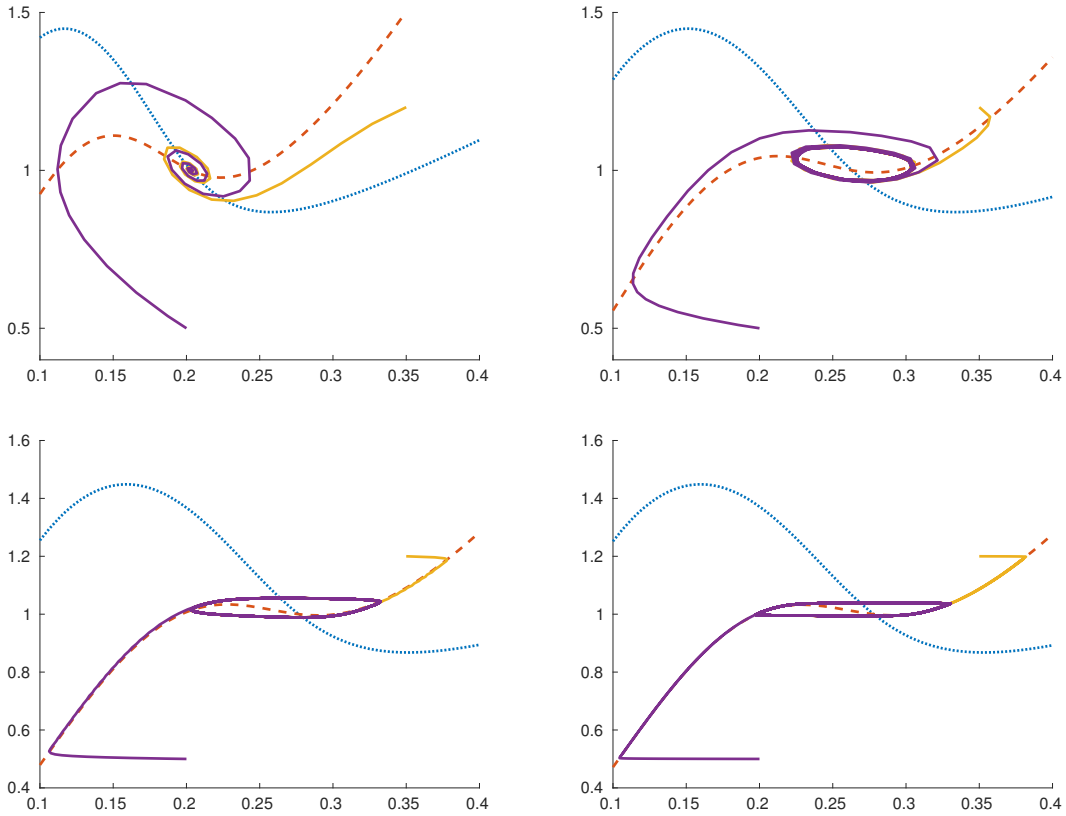



Figure 8.1: $u$ is in $x$-axis, $v$ in $y$-axis. Red dashed curves correspond to nullclines $g_\epsilon = 0$ ($\dot{u_\epsilon} = 0$) and blue dotted curves to nullclines $f_\epsilon = 0$ ($\dot{v_\epsilon} = 0$). The four figures correspond to decreasing values of $\epsilon$ from top-left to bottom-right (0.5, 0.1, 0.01 and 0.001). In yellow and purple, two trajectories $t \mapsto (u_\epsilon(t), v_\epsilon(t))$ are shown, for two different initial conditions (respectively $(0.35, 1.2)$ and $(0.2, 0.5)$).

It turns out that $\overline{L}(\epsilon)\big(\epsilon(b_E - d_L) - \frac{1}{\eta}\big) = d_E$. Hence to guarantee $\eta(\epsilon) = O(1)$ it is required that

$$b_E - d_L = O(1/\epsilon).$$

Therefore the scaling choice made in Theorem 8.1 is in some sense "generic".

Note that for every possible parameter scaling we get a (possibly different) limit in (8.21). For instance, assuming $\epsilon c_\epsilon$ and $\epsilon b_E(\epsilon)$ have limits $1/\eta_0, \xi > 0$ respectively as $\epsilon \to 0$ (this is the case with the scaling used in Theorem 8.1), we choose $\eta(\epsilon) = O(1)$ such that $c_\epsilon \eta(\epsilon)\epsilon = 1$ and end up with

$$\begin{cases} \dot{v} = \epsilon\eta b_E u - \big(d_E + h(\eta u)\big)v =: f_\epsilon(u, v), \\ \epsilon\dot{u} = \frac{1}{\eta}h(\eta u)v - d_L\epsilon u - u^2 =: g_\epsilon(u, v). \end{cases}$$

The limits $f$ and $g$ are given by

$$f(u, v) = \eta_0\xi u - \big(d_E + h(\eta_0 u)\big)v, \quad g(u, v) = \frac{1}{\eta_0}h(\eta_0 u)v - u^2. \tag{8.23}$$

### 8.3.2 Proof of the main result

We proceed to the proof of Theorem 8.1 in three steps. First, scaled quantities $u_\epsilon$ and $v_\epsilon$ remain uniformly bounded independently of $\epsilon$, as can be proved from direct computation using the bound $K$ from Lemma 8.1.

**Lemma 8.5.** *There exists $C > 0$ such that for all $\epsilon > 0$ and $t > 0$,*

$$|u_\epsilon(t)|,\ |v_\epsilon(t)|,\ |f_\epsilon(u_\epsilon(t), v_\epsilon(t))|,\ |g_\epsilon(u_\epsilon(t), v_\epsilon(t))| \le C.$$

Hence, up to extraction, $v_\epsilon$ converges to $v$ uniformly on compact sets $[0,T]$ by the Ascoli theorem. Then, the convergence of an auxiliary quantity gives convergence of $u_\epsilon$:

**Lemma 8.6.** *For all $T>0$.*

$$\|g_\epsilon(u_\epsilon,v_\epsilon)\|_{L^2(0,T)} = O(\sqrt{\epsilon}). \tag{8.24}$$

*Moreover, there exists $u,v \in L^1\cap L^\infty$ such that after extraction of a subsequence $u_\epsilon \to u$ in $L^p(0,T)$ for all $1\le p<\infty$, as $v_\epsilon \to v$ uniformly.*

*Proof.* Let $B(t,u) := \int_0^u g^2(\sigma,v(t))d\sigma$, where $v$ is the limit of $v_\epsilon$ (obtained by the Ascoli theorem) and $g$ is the limit of $g_\epsilon$ (from (8.21)). From (8.22) we deduce that for all $t$, $u\mapsto B(t,u)$ is increasing. Hence there exists a *smooth* function $A(t,u)$ such that for all $t,u$, $A(t,B(t,u))=u$.

If there exists $w(t)\in L^p(0,T)$ for all $p<\infty$ and $T>0$ such that

$$\int_0^{u_\epsilon(t)} g_\epsilon^2(\sigma,v_\epsilon(t))d\sigma \xrightarrow[\epsilon\to 0]{L^p(0,T)} w(t), \tag{8.25}$$

then defining $u(t) := A(t,w(t))$ we can conclude that $u_\epsilon = A\big(\,\cdot\,,\int_0^{u_\epsilon} g^2(\sigma,v)d\sigma\big) \xrightarrow[\epsilon\to 0]{L^p(0,T)} u = A(\,\cdot\,,w)$.

Indeed, we notice that

$$\int_0^{u_\epsilon(t)} g_\epsilon^2(\sigma,v_\epsilon(t))d\sigma - \int_0^{u_\epsilon(t)} g^2(\sigma,v_\epsilon(t))d\sigma \to 0,$$

and

$$\int_0^{u_\epsilon(t)} g^2(\sigma,v_\epsilon(t))d\sigma - \int_0^{u_\epsilon(t)} g^2(\sigma,v(t))d\sigma \to 0.$$

Since $u_\epsilon$ is uniformly bounded,

$$\Big|\int_0^{u_\epsilon(t)} g_\epsilon^2(\sigma,v_\epsilon(t))d\sigma - \int_0^{u_\epsilon(t)} g^2(\sigma,v(t))d\sigma\Big| \le u_\epsilon(t)\big(\|g_\epsilon^2-g^2\|_\infty + C\|v_\epsilon - v\|_\infty\big),$$

for some $C>0$ which depends only on $\partial_v g$. Hence (8.25) implies

$$\int_0^{u_\epsilon(t)} g^2(\sigma,v(t))d\sigma \xrightarrow[\epsilon\to 0]{L^p(0,T)} w(t).$$

Therefore we only need to prove (8.25) to complete the proof. To do so we first obtain (8.24) by computing

$$\begin{aligned}\frac{g_\epsilon(u_\epsilon(t),v_\epsilon(t))^2}{\epsilon} &= g_\epsilon(u_\epsilon(t),v_\epsilon(t))i_\epsilon \\ &= \frac{d}{dt}\int_0^{u_\epsilon(t)} g_\epsilon(\sigma,v_\epsilon(t))d\sigma - f_\epsilon(u_\epsilon,v_\epsilon)\int_0^{u_\epsilon(t)}\partial_v g_\epsilon(\sigma,v_\epsilon(t))d\sigma.\end{aligned}$$

Hence

$$\frac{1}{\epsilon}\int_0^T \big(g_\epsilon(u_\epsilon(t),v_\epsilon(t))\big)^2 dt = \int_{u_\epsilon(0)}^{u_\epsilon(T)} g_\epsilon(\sigma,v_\epsilon(t))d\sigma - \int_0^T f_\epsilon(u_\epsilon(t),v_\epsilon(t))\int_0^{u_\epsilon(t)}\partial_v g_\epsilon(\sigma,v_\epsilon(t))d\sigma dt.$$

Since $f_\epsilon, g_\epsilon$ and $\partial_v g_\epsilon = \frac{1}{\eta(\epsilon)}h(\eta(\epsilon)u_\epsilon)$ are uniformly bounded, we deduce that

$$\int_0^T g_\epsilon(u_\epsilon(t),v_\epsilon(t))^2 dt = O(\epsilon).$$

This gives (8.24). Then we introduce

$$w_\epsilon(t) := \int_0^{u_\epsilon(t)} g_\epsilon^2(\sigma,v_\epsilon(t))d\sigma.$$

We compute

$$\dot{w}_\epsilon(t) = \frac{1}{\epsilon} g_\epsilon^2(u_\epsilon(t), v_\epsilon(t)) \epsilon \dot{i}_\epsilon + f_\epsilon(u_\epsilon(t), v_\epsilon(t)) \int_0^{u_\epsilon(t)} 2 g_\epsilon(\sigma, v_\epsilon(t)) \partial_v g_\epsilon(\sigma, v_\epsilon(t)) d\sigma.$$

By the previous point, $t \mapsto \frac{1}{\epsilon} g_\epsilon^2(u_\epsilon(t), v_\epsilon(t))$ is uniformly (in $\epsilon$) bounded in $L^1$. In addition, $t \mapsto \epsilon \dot{i}_\epsilon(t)$ is uniformly (in $\epsilon$) bounded in $L^\infty$, by the Lemma 8.5. The second term $f_\epsilon \int g_\epsilon \partial_v g_\epsilon$ is uniformly bounded as well.

As a consequence, $w_\epsilon$ is uniformly (in $\epsilon$) bounded in $BV_{\text{loc}}$. This implies that up to extraction, $w_\epsilon \to w$ in $L^1$. Because $w_\epsilon$ is also bounded in $L^\infty$, convergence actually takes place in all $L^p$ spaces.

□

Finally, the shapes of $(f, g)$ allow us to describe simply the limit trajectories. We use the following assumptions: for all $\epsilon > 0$ small enough, we assume that the right-hand sides of system (8.20) satisfy

(R.1) the set $\mathbb{R}^2 \backslash \{f_\epsilon = 0, g_\epsilon = 0\}$ has exactly 4 connected components, whose measures do not vanish as $\epsilon \to 0$,

(R.2) $f(u_0, v_0) \neq 0$, $g(u_0, v_0) \neq 0$ and the couple $\big(sgn(f_\epsilon(u_0^\epsilon, v_0^\epsilon), sgn(g_\epsilon(u_0^\epsilon, v_0^\epsilon)\big)$ is constant and equal to $\big(sgn(f(u_0, v_0)), sgn(g(u_0, v_0))\big)$.

We also assume that the uniform limits $f, g$ of $f_\epsilon, g_\epsilon$ satisfy

(L.1) the curve $\Upsilon := \{g = 0\}$ is the graph of a function $\phi \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R}_+)$ with $\phi(\infty) = \infty$ and $\phi(0) = 0$,

(L.2) the function $g$ is positive on the epigraph of $\phi$,

(L.3) the function $\phi$ has exactly two local extrema,

(L.4) on the graph of $\phi$, $sgn(f) = -1$ except for a bounded set.

**Lemma 8.7.** *With these assumptions we have:*

*There exists a unique $\tau > 0$ and a (unique up to translations) $\tau$-periodic function $(u_\tau, v_\tau) : \mathbb{R}_+ \to \Upsilon$ such that $v_\tau$ is Lipschitz-continuous, $u_\tau$ is piecewise continuous, for all $t \geq 0$, $v_\tau = \phi(u_\tau)$ everywhere, $\dot{v}_\tau = f(u_\tau, v_\tau)$ almost everywhere and the discontinuities of $u_\tau$ are located at times $t$ such that $\phi$ has a local extremum at $u_\tau(t^-)$.*

*There exists $\tau_1 \geq 0$ and $\tau_2 \in [0, \tau)$ such that for all $t > \tau_1$, $(u, v)(t) = (u_\tau, v_\tau)(t + \tau_2)$. Moreover, by construction $\tau_1$ and $\tau_2$ are uniquely defined from $u_0$ and $v_0$, so the limit $(u, v)$ is in fact unique and the whole family $(u_\epsilon, v_\epsilon)_\epsilon$ converges as $\epsilon$ goes to 0.*

Clearly from (8.23), Lemma 8.7 applies with the hypotheses of Theorem 8.1 and

$$\phi(u) = \frac{\eta_0 u^2}{h(\eta_0 u)}, \quad \eta_0 = \frac{\overline{L}^2}{h(\overline{L})}, \tag{8.26}$$

thus proving the remaining part of the theorem.

*Proof of Lemma 8.7.* Thanks to assumptions (R.1), (L.1), (L.3) and (L.4), the construction of $(u_\tau, v_\tau)$ is classical and can be done by pasting together solutions of Cauchy problems given (locally) by $\dot{v}_\tau = f(\phi^{-1}(v_\tau), v_\tau)$, on intervals where $\phi$ is invertible. Uniqueness comes from the crucial fact that discontinuities of $u_\tau$ are assumed to be located at local extrema of $\phi$.

From the previous lemmas we know that $(u, v) \in \Upsilon$ almost everywhere. In addition, uniform boundedness of $f_\epsilon(u_\epsilon, v_\epsilon)$ ensures that $v$ is Lipschitz continuous.

Then, we claim that if $t > 0$ is such that $\phi$ has no local extremum at $u(t^+)$, then there exists $\tau_0 > 0$ such that $(u, v)$ is continuous on $(t, t + \tau_0)$. This point is the key of the proof. To prove it, let $u_i$ be such that $\phi'(u_i) < 0$. We solve only the simpler problem

$$\begin{cases} \dot{\widehat{v}}_\epsilon = f(\widehat{u}_\epsilon, \widehat{v}_\epsilon), \quad \widehat{v}_\epsilon(0) = \phi(u_i) + O(\epsilon), \\ \epsilon \dot{\widehat{u}}_\epsilon = g(\widehat{u}_\epsilon, \widehat{v}_\epsilon), \quad \widehat{u}_\epsilon(0) = u_i + O(\epsilon). \end{cases}$$

Introducing $\widehat{w}_\epsilon := \widehat{u}_\epsilon - \phi^{-1}(\widehat{v}_\epsilon)$, where the inverse of $\phi$ is taken locally (this is possible for $\epsilon$ small enough since $\phi'(u_i) < 0$ and $\widehat{v}_\epsilon$ is uniformly Lipschitz-continuous), we obtain

$$\dot{\widehat{w}}_\epsilon = \frac{\widehat{w}_\epsilon}{\epsilon}\partial_1 g(\widehat{r}_\epsilon(t), \widehat{v}_\epsilon(t)) - \frac{f(\widehat{u}_\epsilon, \widehat{v}_\epsilon)}{\phi'(\phi^{-1}(\widehat{v}_\epsilon))}, \quad \widehat{w}_\epsilon(0) = O(\epsilon),$$

for some $\widehat{r}_\epsilon(t)$ between $\widehat{u}_\epsilon(t)$ and $\phi^{-1}(\widehat{v}_\epsilon(t))$. We have $\partial_1 g \leq -\alpha < 0$ on a neighborhood of $(u_i, \phi(u_i))$, so on this neighborhood $\widehat{w}_\epsilon$ remains small (it is a $o(\epsilon)$), which in turn proves that $(\widehat{r}_\epsilon, \widehat{v}_\epsilon)$ remains in this neighborhood. In particular, $\widehat{u}_\epsilon$ converges to some function $\widehat{u}$ which is continuous at $t = 0$ (since it is equal to $\phi^{-1}(\widehat{v}(t))$ on a positive neighborhood of 0). We do not write the full proof because the derivation we use here extends readily at the price of tedious notations. A full proof should use $f_\epsilon, g_\epsilon$ rather than $f, g$, and rise some analogue $\phi_\epsilon$ of $\phi$ at level $\epsilon > 0$, for $\epsilon$ small enough, which is locally invertible on a neighborhood of the initial data. It does not require more assumptions than the ones we stated.

This is enough to get all the results of Lemma 8.7, except for the initial layer which we treat now. To fix the notations, we assume that $\phi$ has a local minimum equal to $\phi_m$ at $u_m$ and a local maximum equal to $\phi_M > \phi_m$ at $u_M < u_m$. Moreover, let $u_m^0 < u_M$ such that $\phi(u_m^0) = \phi(u_m)$. For $\alpha, \beta \in \{1, -1\}$, we also introduce $Z_\alpha^\beta := \{sgn(f) = \alpha,\ sgn(g) = \beta\}$.

We define a mapping $\pi : \mathbb{R}^2 \to \Upsilon$ by $\pi = Id$ on $\Upsilon$ and if $(u, v) \in Z_\alpha^\beta$ then $\pi(u, v) = (u_1, v)$ such that $\phi(u_1) = v$ and $sgn(u_1 - u) = \beta$. The projection $\pi$ is well-defined thanks to the assumptions on $\phi$ and $g$, except on $(u_m^0, +\infty) \times \{\phi_m\}$, on which we let $\pi \equiv (u_m^0, \phi_m)$. Then $(u, v)(0^+) = \pi(u_0, v_0)$. To prove this, one simply has to check the behavior of $u_\epsilon$ (since $v_\epsilon$ and $v$ are Lipschitz continuous). As above, we claim that the first-order behavior is simply given by the "layer equation"

$$\epsilon\dot{\widetilde{u}}_\epsilon = g(\widetilde{u}_\epsilon, v_0), \quad \widetilde{u}_\epsilon(0) = u_0,$$

which makes $\widetilde{u}_\epsilon$ converge exponentially fast to $\pi(u_0, v_0)_1$, thanks to assumptions (R.2) and (L.2). Up to tedious notations and thanks to (8.21) and (R.2), this result extends to $u_0^\epsilon$, $v_\epsilon$ and $g_\epsilon$.

Let $\Upsilon_u = \Upsilon \cap \big([u_M, u_m] \times \mathbb{R}_+\big)$ and $\Upsilon_s = \Upsilon - \Upsilon_u$. (Note that $\pi(\mathbb{R}_+^2 - \Upsilon_u) = \Upsilon_s$.) After the initial layer, the trajectory of $(u, v)$ remains on $\Upsilon_s$. This follows from the sign of $f$ on $\Upsilon_u$: because of the continuity property, the trajectory cannot exit $\Upsilon_s$ but at $(u_m, \phi_m)$ (or $(u_M, \phi_M)$, respectively). At these points however, $\Upsilon_u$ is repulsive since $v$ must be continuous, $\dot{v} < 0$ ($\dot{v} > 0$, respectively) and $\Upsilon_u$ lies locally in $\{v > \phi_m\}$ (respectively in $\{v < \phi_M\}$).

Still, the initial data does not need to be projected directly by $\pi$ on $\Upsilon_s \cap \mathbb{R}_+ \times [\phi_m, \phi_M]$. Therefore, we introduce $\tau_1 \geq 0$ as

$$\tau_1 := \max\big(0, \sup\{t \geq 0, \quad v(t) \notin [\phi_m, \phi_M]\}\big).$$

It remains to check that $\tau_1 < +\infty$. For all $T > 0$, as long as $\phi$ has no local extremum at $u(t)$ for $t \in (0, T)$, $u$ is continuous. Thanks to our assumption (R.1), there are two connected components in $\Upsilon_s$, on each one of whom $sgn(f)$ is constant. Because of assumption (L.4), $f$ must be negative on the unbounded connected component. Therefore $(u, v)$ remains on $(0, T)$ in a part of $\Upsilon$ where $|f|$ is positively bounded from below (one of the two connected components of $\Upsilon_s$) and has the appropriate sign. This yields the existence of $\tau_1 < +\infty$.

Then for all $t \geq \tau_1$ we have $v(t) \in [\phi_m, \phi_M]$, and the trajectory is uniquely defined onwards. □

**Remark 8.3.** *We did not treat the case when the limit of $(u_0^\epsilon, v_0^\epsilon)$ belongs to $\Upsilon$ (relaxing assumption (R.2)). In this case indeed, no general result can be obtained, unless the various convergence speeds (of $f_\epsilon, g_\epsilon, u_0^\epsilon$ and $v_0^\epsilon$) are quantified.*

**Remark 8.4.** *The last point of Theorem 8.1 implies that the amplitude of the oscillations (in $u, v$) at the limit $\epsilon \to 0$ can be computed if one knows these parameter scale in $\epsilon$ thanks to only $f$ and $g$. Their period $\tau$ can also be computed directly from $f$ and $\phi$. As in the proof of Lemma 8.7 we denote the intervals of values taken by $u(t)$ where it is continuous (and thus $\mathcal{C}^\infty$) as $[u_m^0, u_M]$ and $[u_m, u_M^0]$ respectively, and let*

$$\Psi(u, v) = \int_u^v \frac{\phi'(u')}{f(u', \phi(u'))} du'.$$

*Then we have*

$$\tau = \Psi(u_m^0, u_M) + \Psi(u_M^0, u_m). \tag{8.27}$$

## 8.4 Hopf bifurcation

Numerical observations (see Section 8.2.3 and Appendix 8.A) show that the system (8.1) has a stable periodic solution oscillating around the non-zero steady state, even far from the slow-fast asymptotic we explored in the previous section. We now prove the local existence of this periodic solution using the Hopf bifurcation theorem (Theorem 8.8 from [167], with a classical proof in [162]; see also [90]) for $2 \times 2$ systems of differential equations.

### 8.4.1 The function class $H_{\overline{L}}$

To find out a possible bifurcation parameter, we choose the hatching function $h$ within a special class, for which we fix the value of one specific steady state $\overline{L}$. With this setting, we can state a bifurcation theorem using the simple bifurcation parameter $h'(\overline{L})$, which represents the sensitivity of hatching rate to larval density at equilibrium.

However, it is worth noting that our argument does not rely on the structure of this class of functions, and may be adapted.

For a fixed $\overline{L}$ the class of functions under consideration that fits our purposes is

$$H_{\overline{L}} := \Big\{ h(L) = a\Big( \arctan(b(L-\overline{L})) + \frac{\pi}{2} \Big),\ a, b \in \mathbb{R}^+ \Big\}. \tag{8.28}$$

Graphs of these functions are shown in Figure 8.2. We use the immediate properties that these

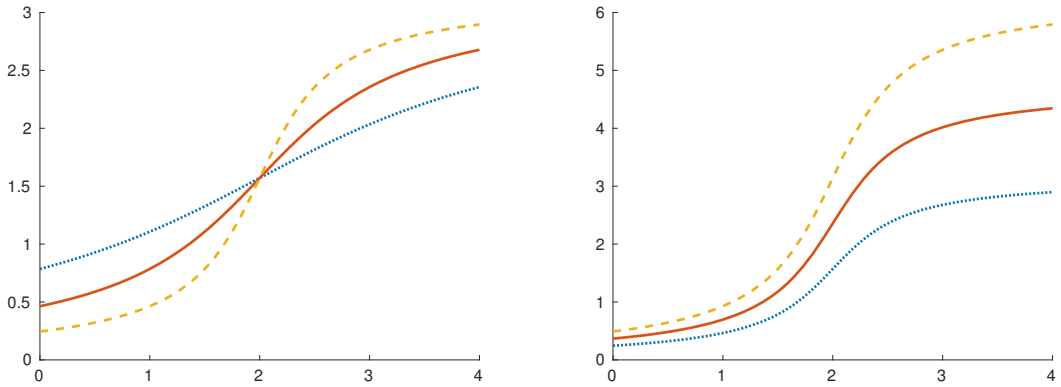



Figure 8.2: Function $h \in H_{\overline{L}}$ with $\overline{L} = 2$. *Left*: $a = 1$, $b = \{0.5, 1, 2\}$. *Right*: $a = \{1, 1.5, 2\}$, $b = 2$. Curve styles with increasing values in $a$ and $b$: dotted blue, solid red, dashed yellow.

functions are positive and increasing. For any couple $(k, k') \in \mathbb{R}_+^* \times \mathbb{R}_+^*$, there exists a unique function $h$ of class $H_{\overline{L}}$ with $h(\overline{L}) = k$ and $h'(\overline{L}) = k'$. Finally, for all $c > 0$, the steady state relation $h(\overline{L}) = \frac{d_E(d_L + c\overline{L})}{b_E - d_L - c\overline{L}}$ has a positive solution in $\overline{L}$ if $a > \frac{d_E d_L}{b_E - d_L} \frac{2}{\pi}$. Indeed, for given values $(k, k') \in \mathbb{R}_+^2$, the choice of $a = \frac{2k}{\pi}$ and $b = \frac{k'}{a}$ gives the solution since

$$h(\overline{L}) = a\frac{\pi}{2} = \frac{2k}{\pi}\frac{\pi}{2} = k \text{ and } h'(\overline{L}) = ab = \frac{2k}{\pi}\frac{k'}{\frac{2k}{\pi}} = k'.$$

Also we can solve the equation in $\overline{L}$, $\frac{a\pi}{2} = \frac{d_E(d_L + c\overline{L})}{b_E - d_L - c\overline{L}}$, which yields $\overline{L} = \frac{\frac{a\pi}{2}(b_E - d_L) - d_E d_L}{c\frac{a\pi}{2} + c d_E}$. Hence $\overline{L}$ is positive under the stated condition.

**Remark 8.5.** *From Lemma 8.2, for $h$ of class $H_{\overline{L}}$, the state $(0,0)$ is unstable if and only if*

$$a > \frac{d_E d_L}{(b_E - d_L)(\frac{\pi}{2} + \arctan(-b\overline{L}))}.$$

### 8.4.2 Transformation into a canonical form

Let $P = (a, b) \in \mathbb{R}_+^2$ and the function $h_P$ of class $H_{\overline{L}}$

$$h_P(L) = a\Big( \arctan(b(L-\overline{L})) + \frac{\pi}{2} \Big). \tag{8.29}$$

We use the notation $k := h_P(\overline{L}) = a\frac{\pi}{2}$. Let $P : \gamma \mapsto P(\gamma) = (a_0, b_0 + \gamma)$ where $(a_0, b_0) \in \mathbb{R}_+^{*^2}$. Then we can associate $P(\gamma)$ to a new system $(S_\gamma(a_0, b_0))$ obtained from (8.1)

$$\begin{cases} \dot{E} = b_E L - d_E E - h_{P(\gamma)}(L)E, \\ \dot{L} = h_{P(\gamma)}(L)E - d_L L - cL^2. \end{cases} \qquad (S_\gamma(a_0, b_0))$$

This system has a positive equilibrium $(\overline{E}, \overline{L})$ and the Jacobian matrix of the system evaluated in $(\overline{E}, \overline{L})$ is:

$$J_{P(\gamma)} = \begin{pmatrix} -d_E - h_{P(\gamma)}(\overline{L}) & b_E - h'_{P(\gamma)}(\overline{L})\overline{E} \\ h_{P(\gamma)}(\overline{L}) & h'_{P(\gamma)}(\overline{L})\overline{E} - d_L - 2c\overline{L} \end{pmatrix},$$

We set $\lambda_{1,2}(\gamma) = \alpha(\gamma) \pm i\beta(\gamma)$ the eigenvalues of $J_{P(\gamma)}$, when the discriminant of the characteristic polynomial of $J_{P(\gamma)}$ is negative.

### 8.4.3 Main result

Using function $T$ from (8.12), we define

$$b(a) := \frac{T(a)}{a}, \quad a_{crit} := \frac{2k^+}{\pi} > 0. \qquad (8.30)$$

**Theorem 8.2.** *There exists $\tilde{a} > 0$ such that: If $a > max(\tilde{a}, a_{crit})$, $(S_\gamma(a, b(a)))$ has a supercritical Hopf Bifurcation in $\gamma = 0$. In particular:*

1. *there exists $\gamma_1$¡0 such that for all $\gamma \in (\gamma_1, 0]$, $(\overline{E}, \overline{L})$ is a stable focus,*
2. *for all $U$ neighborhood of $(\overline{E}, \overline{L})$, there exists $\gamma_2 > 0$ such that for all $\gamma \in [0, \gamma_2)$, $(\overline{E}, \overline{L})$ is an unstable focus surrounded by a stable limit cycle contained in $U$, which has an amplitude that grows when $\gamma$ grows.*

**Remark 8.6.** *$k_+$ is given by (8.13), and $\tilde{a}$ is such that the normal form coefficient $\alpha_N$ (see [167]) of our system is negative if $a > \tilde{a}$. We simply give a numerical justification of the existence of $\tilde{a}$ as the computations appear to be very long (see the proof below).*

**Remark 8.7.** *The value of $a$ must be greater than $a_{crit}$ to ensure that the linearized operator has complex eigenvalues.*

The bifurcation diagram for $S_\gamma(a_0, b_0)$ in Figure 8.3 is obtained by XPPAUT software [80].

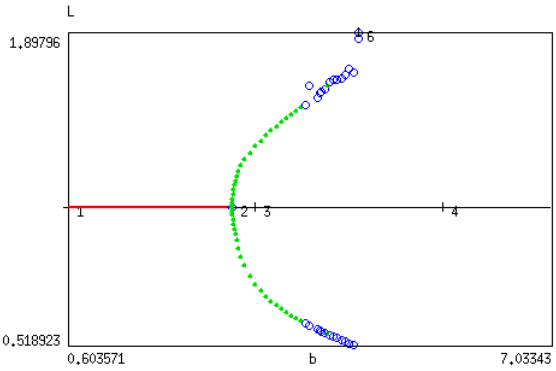



Figure 8.3: Supercritical Hopf bifurcation diagram with $a_0 = 0.2$. The bifurcation parameter $b$ is in $x$-axis, the diagram shows extreme values of the periodic solution for $L$ (the $L$ scale is in $y$-axis). The steady state is stable (red line) until the bifurcation point (point number 2) is reached. A periodic solution appears and is stable (green points) until a bigger value of $b$, where it becomes unstable (blue circles). The amplitude of the periodic solution grows with the parameter $b$.

*Proof of Theorem 8.2.* We set $\lambda_{1,2}(\gamma) = \alpha(\gamma) \pm i\beta(\gamma)$ (with $\gamma$ a real parameter), the two eigenvalues of $J_{P(\gamma)}$ the Jacobian matrix associated to our system and computed in $(0,0)$. We call $\gamma_c$ a bifurcation value, and $\alpha_N(\gamma)$ the normal form coefficient of the system (see [167]).

Firstly, we only need to study complex conjugate and pure imaginary eigenvalues of $J_{P(\gamma)}$ to find the bifurcation value $\gamma_c$, which means also to look for $\gamma_c$ such that $\alpha(\gamma_c) = 0$ and $\beta(\gamma_c) \neq 0$. Thanks to Proposition 8.1 we know that this is the case when $k > k_+$ *i.e.* $\frac{a\pi}{2} > k_+$ or equivalently $a > a_{crit}$ (by definition, $a_{crit} = \frac{2k_+}{\pi}$). Moreover since $h'(\overline{L}) = ab$ (direct computation from (8.29)), we know that the bifurcation value is located at the level of the graph $G$ of function $b$ defined in (8.30)

$$G := \{(a,b) \in \mathbb{R}^2,\ a > a_{crit},\ T(a) = ab = h'_{P(\gamma)}(\overline{L})\}. \tag{8.31}$$

And we can set $\gamma_c = 0$.

Secondly, we have to see if $\frac{d\alpha}{d\gamma}(\gamma_c) > 0$, this means to check that $\mathrm{tr}(J_{P(\gamma)})$ changes sign at the bifurcation value $\gamma_c$. Let $\gamma \longmapsto z(\gamma) = \alpha(a_0, b(a_0) + \gamma)$. We recall that $\alpha(\gamma)$ depends on $a$ and $b$.

Since

$$\alpha = \frac{\mathrm{tr}(J_{P(\gamma)})}{2} = \frac{1}{2}\Big( - d_E - \frac{a\pi}{2} + ab\overline{E} - d_L - 2c\overline{L}\Big),$$

we have $z'(\gamma) = \partial_b \alpha = \frac{a\overline{E}}{2}$ and we obtain that $z'(\gamma_c) = z'(0) = \frac{a\overline{E}}{2}$ and it is always positive.

Thirdly, we have to study the normal form coefficient of the system computed in $\gamma_c = 0$ and find when $\alpha_N(\gamma_c) \neq 0$. To get the normal form coefficient, we have to transform the system $(S_\gamma(a_0, b_0))$ and we use the steps from [167]. In a first step we reduce the initial system $(S_\gamma(a_0, b_0))$ to a system where the equilibrium $(\overline{E}, \overline{L})$ becomes the origin. By the change of variables $x = E - \overline{E}$ and $y = L - \overline{L}$, $(S_\gamma(a_0, b_0))$ becomes:

$$\begin{cases} \dot{x} = b_E(y + \overline{L}) - d_E(x + \overline{E}) - a\Big(\arctan(by) + \dfrac{\pi}{2}\Big)(x + \overline{E}), \\ \dot{y} = a\Big(\arctan(by) + \dfrac{\pi}{2}\Big)(x + \overline{E}) - d_L(y + \overline{L}) - c(y + \overline{L})^2. \end{cases} \tag{8.32}$$

Then as $(\overline{E}, \overline{L})$ is an equilibrium, we can simplify (8.32) into

$$\begin{cases} \dot{x} = b_E y - d_E x - \dfrac{a\pi}{2}x - a\Big(\arctan(by)\Big)(x + \overline{E}), \\ \dot{y} = a\Big(\arctan(by) + \dfrac{\pi}{2}\Big)(x + \overline{E}) + \dfrac{a\pi}{2}x - d_L y - cy^2 - 2cy\overline{L}, \end{cases} \tag{8.33}$$

which we write as

$$\begin{cases} \dot{x} = b_E y - d_E x - \dfrac{a\pi}{2}x - aby\overline{E} + f(x,y), \\ \dot{y} = \dfrac{a\pi}{2}x - d_L y - 2cy\overline{L} + aby\overline{E} + g(x,y), \end{cases} \tag{8.34}$$

where

$$\begin{aligned} f(x,y) &= aby\overline{E} - a\arctan(by)(x + \overline{E}), \\ g(x,y) &= -aby\overline{E} + a\arctan(by)(x + \overline{E}) - cy^2. \end{aligned}$$

The system (8.34) can also be written under the matrix form

$$\begin{pmatrix} \dot{x} \\ \\ \dot{y} \end{pmatrix} = \begin{pmatrix} -\frac{a\pi}{2} - d_E & b_E - ab\overline{E} \\ \\ \frac{a\pi}{2} & -d_L - 2c\overline{L} + ab\overline{E} \end{pmatrix} \begin{pmatrix} x \\ \\ y \end{pmatrix} + \begin{pmatrix} f(x,y) \\ \\ g(x,y) \end{pmatrix}.$$

We call $M$ the first $(2 \times 2)$ matrix in the right-hand-side.

Now, to obtain the normal form coefficient, one way is to perform a linear change of variables so as to get

$$\begin{pmatrix} \dot{X} \\ \\ \dot{Y} \end{pmatrix} = N \begin{pmatrix} X \\ \\ Y \end{pmatrix} + \begin{pmatrix} F(X,Y) \\ \\ G(X,Y) \end{pmatrix}, \quad N := \begin{pmatrix} 0 & -\omega \\ \\ \omega & 0 \end{pmatrix}. \tag{8.35}$$

In our case, we can have an idea of the normal coefficient only in a neighborhood of $\gamma = 0$. Because we want to make a simple linear change of variables, we are looking for a matrix $P$ such

that $PMP^{-1} = N$ and that *at the bifurcation value* $\gamma = 0$, $\mathrm{tr}(M) = 0 = \mathrm{tr}(N)$ and $\det(M) = -A^2 - BC = \omega^2 > 0$.

We set $M = \begin{pmatrix} A & B \\ C & -A \end{pmatrix}$ and we can choose $P = \begin{pmatrix} \frac{\omega+A}{2B\omega} & \frac{1}{2\omega} \\ \frac{\omega-A}{2B\omega} & -\frac{1}{2\omega} \end{pmatrix}$, $P^{-1} = \begin{pmatrix} B & B \\ \omega - A & -A-\omega \end{pmatrix}$.

Next we obtain the matrix system (8.35) where

$$\begin{pmatrix} X \\ Y \end{pmatrix} = P \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{x(\omega+A)}{2B\omega} + \frac{y}{2\omega} \\ \frac{x(\omega-A)}{2B\omega} - \frac{y}{2\omega} \end{pmatrix}, \quad \begin{pmatrix} x \\ y \end{pmatrix} = P^{-1} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} (X+Y)B \\ (\omega - A)X + (-\omega - A)Y \end{pmatrix},$$

$$\begin{pmatrix} F(X,Y) \\ G(X,Y) \end{pmatrix} = P \begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix} = \begin{pmatrix} \frac{\omega+A}{2B\omega} f(x,y) + \frac{1}{2\omega} g(x,y) \\ \frac{\omega-A}{2B\omega} f(x,y) - \frac{1}{2\omega} g(x,y) \end{pmatrix} = \begin{pmatrix} f(x,y)\left(\frac{\omega+A}{2B\omega} - \frac{1}{2\omega}\right) - \frac{1}{2\omega} cy^2 \\ f(x,y)\left(\frac{\omega-A}{2B\omega} + \frac{1}{2\omega}\right) + \frac{1}{2\omega} cy^2 \end{pmatrix}.$$

In a final step we compute the normal form coefficient using the previous formulas and the expression that exists in two dimensions given in [167] which is:

$$\begin{aligned} \alpha_N(\gamma = 0) \quad = \quad & \frac{1}{16}\Big(F_{XXX} + F_{XYY} + G_{XXY} + G_{YYY}\Big) \\ & -\frac{1}{16\omega}\Big(G_{XY}(G_{XX} + G_{YY}) - F_{XY}(F_{XX} + F_{YY}) + F_{XX}G_{XX} - F_{YY}G_{YY}\Big). \end{aligned}$$

The coefficient is easy but very tedious to compute, and we used the computer algebra system Maple [1] to get its expression.

In our case the coefficient is equal to zero for some value $\tilde{a} > 0$, and is always negative for $a > \tilde{a}$ (as it appears that $\tilde{a} < a_{crit}$, this is sufficient by definition of (8.31)). Then $\alpha_N(\gamma_c) \neq 0$ for $a \neq \tilde{a}$.

Finally, we want to have for all real $\gamma$ in a neighborhood of 0, $\alpha_N(\gamma)\alpha(\gamma) < 0$. Thanks to Maple we have $\alpha_N(0) < 0$, in a neighborhood of $\gamma = 0$, for $a > \tilde{a}$ with $\tilde{a}$ small.

So we can apply the Hopf bifurcation theorem that ensures there exists a limit cycle (periodic solution) when $\alpha(\gamma) > 0$ (i.e $\mathrm{tr}(J_P(\gamma)) > 0$), and moreover this cycle is stable as $\alpha(\gamma) > 0$: we are faced to a supercritical bifurcation. □

### 8.4.4 Discussion on the period of the oscillations

Another point is the study of periods of these solutions because they can be compared with observations in nature.

**Proposition 8.3.** *As $\gamma \to 0^+$, the periodic solution of the system $(S_\gamma(a, b(a)))$ has a frequency $\omega$ and a period $T_0 = 2\pi/\omega$ given by the expression*

$$\omega = \frac{1}{\sqrt{d_E + k}}\Big[k^2(b_E - d_L - d_E) + k(-2{d_E}^2 - b_E d_E - d_E d_L) - {d_E}^3\Big]^{\frac{1}{2}}.$$

*Proof.* As $\gamma \to 0^+$, the oscillations frequency is given by the imaginary part of the root of the polynomial equation (8.9) in the case of non-trivial steady state. The frequency is $\omega_\gamma = \sqrt{\det(J_{P(\gamma)})}$, where the expression of $\det(J_{P(\gamma)})$ is

$$\det(J_{P(\gamma)}) \quad = \quad \frac{1}{d_E + k}\Big[k^2(b_E - d_L - d_E) + k(-2{d_E}^2 - b_E d_E - d_E d_L) - {d_E}^3\Big].$$

Then the expression of $\omega = \omega_0$ follows. □

**Remark 8.8.** *At the bifurcation value, the parameter $k$ can be linked with the period $T_0$. Let $T_0$ a given period observed experimentally, then $k$ is the positive root of the following characteristic polynomial:*

$$k^2\Big[T_0^2(b_E - d_L - d_E)\Big] + k\Big[T_0^2(-2d_E^2 - b_E d_E - d_E d_L) - 4\pi^2\Big] - T_0^2 d_E^3 - 4\pi^2 d_E.$$

*Away from the bifurcation value, the real part of the eigenvalues is greater than zero and the period of the oscillations can only be obtain numerically. Unfortunately, this case is more relevant as the Hopf bifurcation theorem asserts that the amplitude is increasing with the parameter $\gamma$. In other words, for fixed $a$ the amplitude of the oscillations is an increasing function of $b$.*

## 8.5 Conclusion

We show that introducing internal regulation in the form of a larval-density-mediated hatching rate in a compartmental model for mosquito population dynamics induces stable oscillations. These oscillations can be rather simply understood from the mathematical point of view either as cycles produced by a Hopf bifurcation (Theorem 8.2), in a first parameter regime, or as the typical slow-fast behavior (close to FitzHugh-Nagumo model, Theorem 8.1) in a second parameter regime.

Our study supports the idea that understanding internal life-cycle regulation can effectively help modeling and simulating population dynamics properly. Ongoing experiments of some of the authors try to reproduce the larval density impact on hatching which was observed in [78] and may shed some light on this misunderstood phenomenon. In particular, restricting the parameters and possible oscillations range could only be reached by assessing as precisely as possible the actual hatching feedback.

A limitation of our strategy is that it neglects environmental variations. Therefore it leaves open for future studies the deep question of linking internal life-cycle regulation and external variations (induced, for instance, by rainfall and temperature) in order to get a better description of the mosquito populations dynamics. However, it was observed that population oscillations may happen on periods much shorter than seasonal variations, and this justifies the study of internal regulations as possible triggers.

Another possible extension of our works is the adaptive dynamics of hatching regulation trait. Indeed, synchronizing the egg hatching may be beneficial for a population in a given environment, but also be detrimental if rare and extreme events can annihilate larval population, for instance. The egg stage can be seen indeed as a quiescent, refuge state for the species (this approached was studied in [237]). Here we prove that positive feedback of larvae on egg hatching tends to make the population size oscillate, creating distinct generations (*synchronizing effect*) while negative feedback tends to stabilize the population size, which may be detrimental on the long run if, for example, the favorable period for larvae and adult development is typically short.

**Acknowledgements.** BP has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 740623). MS, NV and DAM acknowledge partial funding from Inria, France and CAPES, Brazil (processo 99999.007551/2015-00), in the framework of the STIC AmSud project MOSTICAW and from CAPES/COFECUB project Ma-833 15 "Modeling innovative control method for Dengue fever". MS and NV acknowledge partial funding from the ANR blanche project Kibord: ANR-13-BS01-0004 funded by the French Ministry of Research.

# Appendices

## 8.A Numerical tests for hatching rate given by Hill functions

To explore the possible behaviors depending on the function $h$ of type (8.16), we fix the biological parameters (including $\overline{L}$), $p > 1$ and $k = h(\overline{L}) > 0$ such that (8.19) holds. We introduce the notation $X(k) := \frac{1}{p}(2 + \frac{k-d_L}{b_E}) < 1$ and use two parameters: $\iota \in (0, 1 - X(k))$ and $\zeta \in (0, 1)$, in order to represent the full range of (8.18). More precisely, we will parametrize $a$ and $\alpha$ with $\iota, \zeta$, as functions of $k$, and then we can go back to a function in (8.16) by letting $\lambda = \alpha\overline{L}$ and $h_m = k - a/(1 + \alpha^p)$.

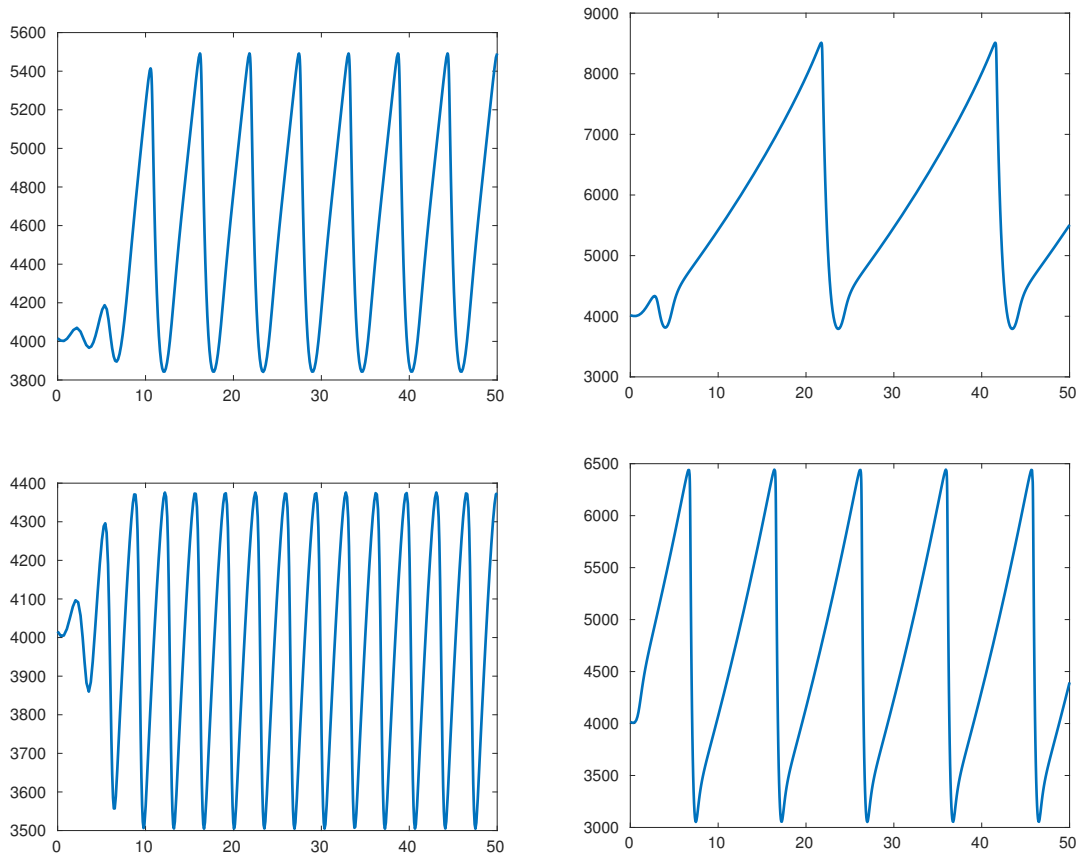



Figure 8.A.1: Egg dynamics from (8.1) for $h$ of Hill function type. All parameters being fixed, including $p = 3$ and $k = 0.5$, $\iota = 0.05$ (top) or $\iota = 0.2$ (bottom) and $\zeta = 0.2$ (left) or $\zeta = 0.8$ (right).

We choose

$$\alpha_\iota(k) = \Big(\frac{X(k) + \iota}{1 - \big(X(k) + \iota\big)}\Big)^{1/p}$$

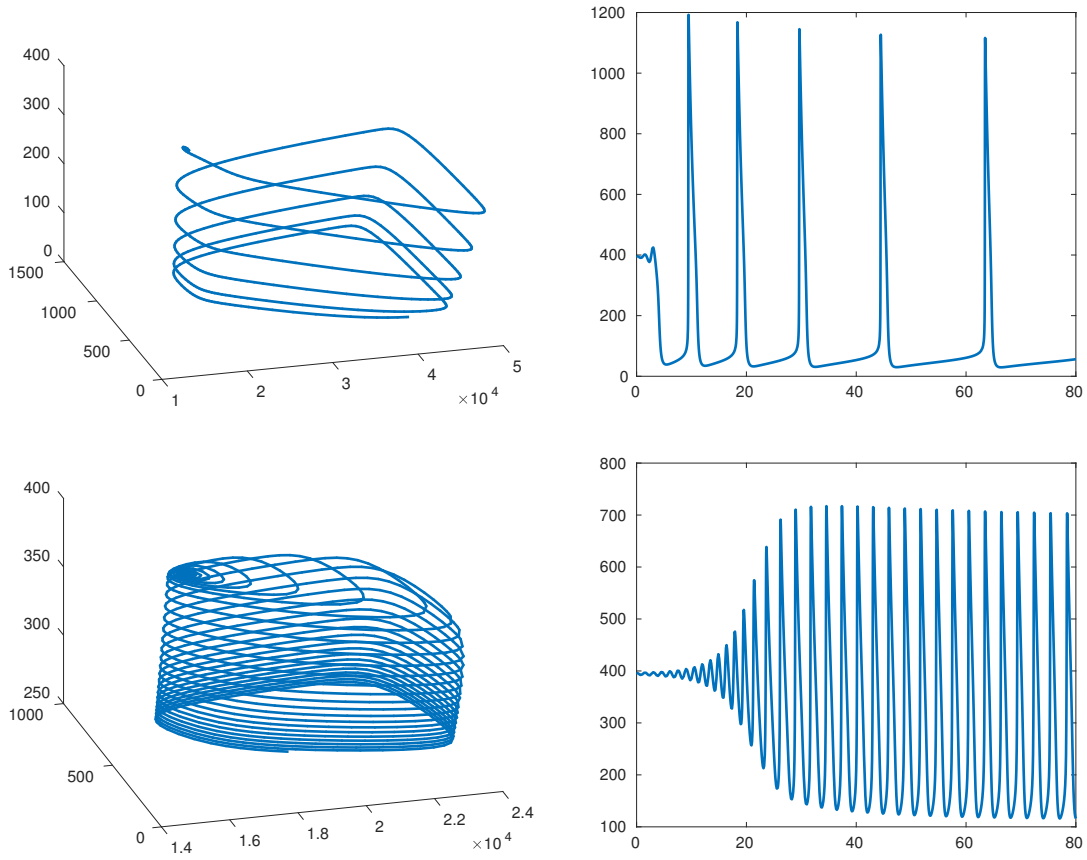



Figure 8.A.2: Numerical solutions of ($S_3$) for $h$ defined by two different Hill functions. All parameters being fixed, including $p = 3$, $k = 0.5$, $\iota = \frac{1-X(k)}{10}$ and choosing $\zeta = 0.1$ (top) or $\zeta = 0.9$ (bottom).

and

$$a_{\zeta,\iota}(k) = \zeta k(1+\alpha_\iota^p) + (1-\zeta)k\frac{(1+\alpha_\iota^p)^2}{\alpha_\iota^p}X(k)$$
$$= k\frac{1-\zeta(1-(X(k)+\iota))}{(X(k)+\iota)(1-(X(k)+\iota))}.$$

For any choice of $\iota$ and $\zeta$, we end up with system (8.1), $h$ given by (8.16), featuring a unique, (locally linearly) unstable positive steady state. At least numerically, solutions always exhibit periodic oscillations, as can be seen in Figure 8.A.1 for egg dynamics.

The above computations extend to the full, 3-dimensional system ($S_3$), and numerical observations are very similar. Indeed, the condition (8.19) guaranteeing positivity of the trace of the Jacobian at the unique positive equilibrium, rewrites for system ($S_3$) as

$$(p-2)\frac{\beta_E\tau_L}{\delta_A} + \delta_L + \tau_L - \delta_A > k.$$

In this case we define

$$X(k) := \frac{\delta_A}{p\beta_E\tau_L}\big(2\frac{\beta_E}{\delta_A}\tau_L - \delta_L - \tau_L + \delta_A + k\big),$$

and the above condition is equivalent to $X(k) < 1$.

Exactly as in the two-dimensional case, we explore the full range of (8.18) by choosing the parameters $(\iota,\zeta) \in (0, 1-X(k)) \times (0,1)$ and defining $\alpha_\iota(k)$ and $a_{\iota,\zeta}(k)$ by the same formulas as before. For all the numerical values we took for $\iota$ and $\zeta$, we always found oscillating solutions. Examples (dynamics of larvae and of $(E, L, A)$ in the three dimensional space) are shown in Figure 8.A.2.

## 8.B Amplitude and period computation in slow-fast regime

In the slow-fast approach, system (8.1) exhibits oscillations with known amplitude and period at the limit $\epsilon \to 0$. We show here how to compute this amplitude analytically. To do so, we simply compute the local extrema of $u \mapsto \frac{\eta u^2}{h(\eta u)}$. The first-order necessary condition yields

$$xh'(x) = 2h(x), \quad x = \eta u.$$

This provides with a general method to determine the limit trajectories. With the previous example from (8.16), $h(x) = h_m + a\dfrac{x^p}{(\alpha\overline{L})^p + x^p}$, this boils down to

$$2(h_m + a)x^{2p} + (\alpha\overline{L})^p\big((2-p)a + 4h_m\big)x^p + 2h_m(\alpha\overline{L})^{2p} = 0.$$

Letting $y = x^p$, we end up with a second-order polynomial, for which the analytical computation can be pushed a few steps further. In particular, its discriminant is

$$\begin{aligned}\Delta &= (\alpha\overline{L})^{2p}\Big(\big((2-p)a + 4h_m\big)^2 - 16h_m(h_m + a)\Big)\\ &= (\alpha\overline{L})^{2p}a\big((2-p)^2a - 8ph_m\big).\end{aligned}$$

Hence there are exactly two positive local extrema if and only if

$$(2-p)^2a > 8ph_m \quad \text{and} \quad (2-p)a + 4h_m < 0.$$

The first condition implies the second one if $p > 2$, and the second one is impossible if $p \leq 2$. Therefore the only case when there are two local extrema is when $p > 2$ and

$$\frac{h_m}{a} < \frac{(p-2)^2}{8p}. \tag{8.36}$$

Under assumption (8.36) we find that the extrema ($y_M < y_m$) are located at

$$(\alpha\overline{L})^p\frac{(p-2)a - 4h_m \pm \sqrt{a^2(p-2)^2 - 8aph_m}}{4(h_m + a)}.$$

Let $\xi_\pm = (p-2)a - 4h_m \pm \sqrt{a^2(p-2)^2 - 8aph_m}$. With the notations of Lemma 8.7,

$$\begin{aligned}u_m &= \frac{\alpha}{\eta}\overline{L}\Big(\frac{\xi_+}{4(h_m+a)}\Big)^{1/p}, \quad \phi_m = \frac{\eta u_m^2}{h(\eta u_m)},\\ u_M &= \frac{\alpha}{\eta}\overline{L}\Big(\frac{\xi_-}{4(h_m+a)}\Big)^{1/p}, \quad \phi_M = \frac{\eta u_M^2}{h(\eta u_M)}.\end{aligned}$$

Then we can compute $u_r^0$ for $r \in \{m, M\}$ by solving $\frac{\eta\cdot(u_r^0)^2}{h(\eta u_r^0)} = \phi_r$. Unfortunately this cannot be done analytically. However, the amplitude of the oscillations in terms of $v$ is equal to

$$A_v := \phi_M - \phi_m.$$

With $\overline{E} = 1/\epsilon$, we expect that the oscillations of $E$ have amplitude

$$\frac{\phi_M - \phi_m}{\epsilon} = \frac{\alpha^2\overline{L}^2}{\eta\epsilon}\Big(\frac{\big(\frac{\xi_-}{4(h_m+a)}\big)^{2/p}}{h_m + a\frac{\xi_-}{h_m+\xi_-}} - \frac{\big(\frac{\xi_+}{4(h_m+a)}\big)^{2/p}}{h_m + a\frac{\xi_+}{h_m+\xi_+}}\Big),$$

where $\eta = \dfrac{\overline{L}^2}{h(\overline{L})} = \dfrac{\overline{L}^2}{h_m + \frac{a}{1+\alpha^p}}$, by (8.26). Hence the amplitude of egg oscillations is equal to

$$\frac{1}{\epsilon}A_v = \frac{1}{\epsilon}\frac{\alpha^2(h_m + \frac{a}{1+\alpha^p})}{(4(h_m+a))^{2/p}}\Big(\frac{\xi_-^{2/p}}{h_m + a\frac{\xi_-}{h_m+\xi_-}} - \frac{\xi_+^{2/p}}{h_m + a\frac{\xi_+}{h_m+\xi_+}}\Big).$$

We can simplify this expression one step further by letting $\rho := h_m/a$. Then we notice that $q_\pm := \xi_\pm/a = p-2-4\rho \pm \sqrt{(p-2)^2-8p\rho}$ and deduce

$$A_v = \frac{\alpha^2}{1+\alpha^p}\frac{1+\rho+\alpha^p}{(4(1+\rho))^{2/p}}\Big(\frac{(\rho q_-)^{2/p}}{1+\frac{\rho^2 q_-}{1+\rho q_-}} - \frac{(\rho q_+)^{2/p}}{1+\frac{\rho^2 q_+}{1+\rho q_+}}\Big).$$

In particular we notice that the amplitude depends only on the function $h$ through $\rho$, $\alpha$ (hence $\overline{L}$) and $p$, and not on any other biological parameter, under the constraints

$$p>2, \quad \rho < \frac{(p-2)^2}{8p}.$$

An interesting case is when $p \to +\infty$, where $h$ approaches a step function from $h_m$ to $h_m+a$, with its jump located at $\alpha\overline{L}$. In this limit we can compute the amplitudes in $u$ and $v$:

$$\begin{cases} A_u = \frac{\alpha}{\overline{L}}(h_m + a\mathbb{1}_{\alpha<1} + \frac{a}{2}\delta_{\alpha=1})\big(\sqrt{\frac{\rho+1}{\rho}} - \sqrt{\frac{\rho}{\rho+1}}\big), \\ A_v = \alpha^2(h_m + a\mathbb{1}_{\alpha<1} + \frac{a}{2}\delta_{\alpha=1})\frac{1}{\rho(h_m+a)}. \end{cases}$$

Using formula (8.27), we can also obtain in this case an analytical expression for the period of the oscillations:

$$\tau = \frac{2}{h_m}\log\Big(\frac{h_m+a/2-\alpha\overline{L}}{h_m+a/2-\alpha\overline{L}\sqrt{\frac{\rho}{1+\rho}}}\Big) + \frac{2}{h_m+a}\log\Big(\frac{h_m+a/2-\alpha\overline{L}\sqrt{\frac{1+\rho}{\rho}}}{h_m+a/2-\alpha\overline{L}}\Big)$$

Indeed, $h(u) = h_m$ if $u < \alpha\overline{L}$ and $h(u) = h_m + a$ if $u > \alpha\overline{L}$ so that if we assume $d_E = 0$ (for simplicity), we get $f(u,\phi(u)) = \eta_0 u\big(\xi - u\big)$ and $\phi'(u) = \frac{2\eta_0}{h_m}u$ if $u < \alpha\overline{L}$ and $\phi'(u) = \frac{2\eta_0}{h_m+a}u$ if $u > \alpha\overline{L}$.

## 8.C Numerical oscillations, period and amplitude close to the bifurcation

We illustrate the statements from Section 8.4 with numerical examples. Biological parameters of (8.1) are taken at a temperature around $25°C$ which leads to $\overline{A} = 3.4$ mosquitoes per 100 square meters (taken from a physical situation described in [229]) and $b_E = 20.94, d_L = 0.15$. (taken from [238]). To fit the condition $d_E \ll d_L$, $d_E$ is fixed arbitrarily at $\frac{1}{180}$. We note that condition (8.3) is satisfied: $b_E = 20.94 > 0.15 + \frac{1}{180} = d_L + d_E$.

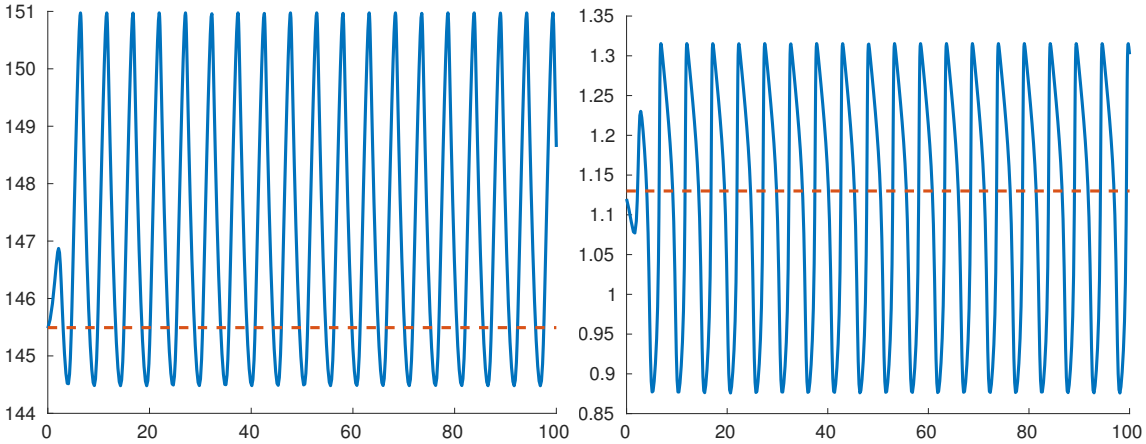



Figure 8.C.1: Time dynamics of eggs (left) and larvae (right) for $a_1 = 0.1$, $b_1^{0.05} = 2.91$.

The parameters $a$ and $b$ are chosen so that Theorem 8.2 applies, which proves the existence of periodic solutions close to the non-trivial steady states. We perform numerical test by letting

a parameter $j$ vary in a set $J$ of 18 values between 0.05 and 4 in order to obtain 162 couples $(a_i, b_i^j)_{i=1,\dots,9;j\in J}$ by

$$a_i = 0.1 + 0.05(i-1) \text{ and } b_i^j = b_{i,min} + j \times b_{i,min},$$

where $b_{i,min}$ is the minimal $b$ that can be chosen for $a_i$ to obtain oscillations (if $b < b_{i,min}$ the solutions can not oscillate), *i.e.* for which the trace of the linearized operator is equal to 0.

The hatching functions are:

$$h_i^j(L) = a_i\Big(\arctan(b_i^j(L-\overline{L})) + \frac{\pi}{2}\Big).$$

In our tests the steady state changes with $i$ (for example $\overline{E}_1 = 145.92$, $\overline{E}_4 = 59.59$ and $\overline{E}_9 = 30$) but we always have $\overline{L} = \overline{A}\frac{\delta_A}{\tau_L} = 1.13$.

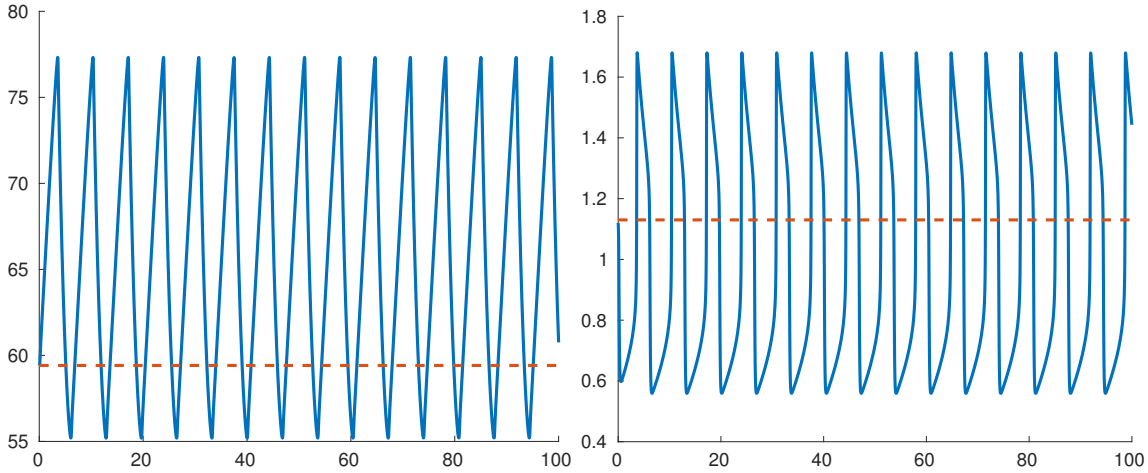



Figure 8.C.2: Time dynamics of eggs (left) and larvae (right) for $a_2 = 0.25$, $b_2^{0.5} = 4.18$.

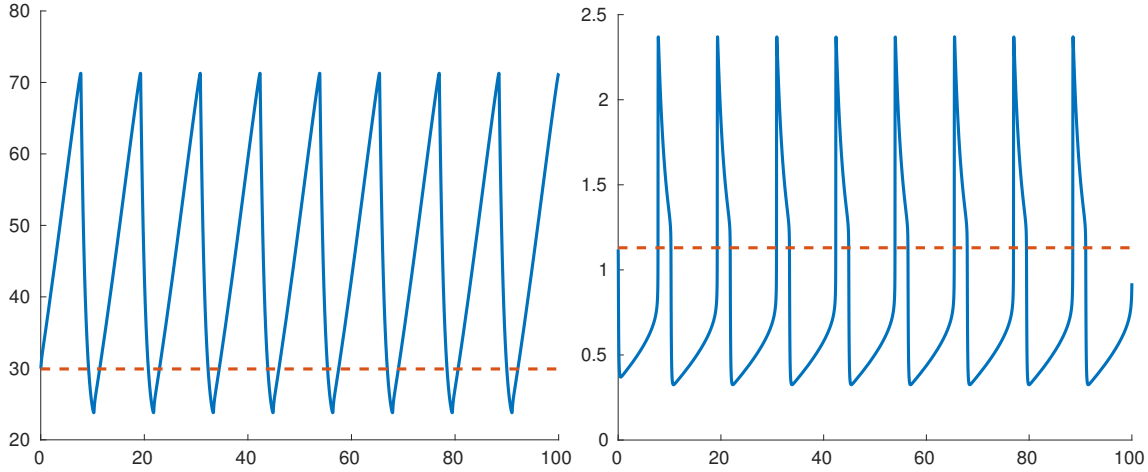



Figure 8.C.3: Time dynamics of eggs (left) and larvae (right) for $a_3 = 0.5$, $b_3^2 = 8.44$.

| $a = 0.1$ | Period (days) | $\overline{E}$ | $\overline{L}$ | Larvae amplitude (%$\overline{L}$) |
|---|---|---|---|---|
| $b = 2.91$ | 5.18 | 145.92 | 1.13 | 23.1 |
| $b = 4.16$ | 15.06 | | | 51.3 |
| $b = 8.32$ | 61.54 | | | 110.22 |
| $b = 3.52$ | 9.6 | | | 42.08 |

Table 8.C.1: Steady states, period and amplitude of oscillations for $a = .1$

| $a = 0.25$ | Period (days) | $\overline{E}$ | $\overline{L}$ | Larvae amplitude ($\%\overline{L}$) |
|---|---|---|---|---|
| $b = 2.93$ | 2.68 | 59.59 | 1.13 | 20.35 |
| $b = 4.18$ | 6.96 | | | 50.67 |
| $b = 8.37$ | 22.64 | | | 110.2 |
| $b = 4.99$ | 9.98 | | | 63.03 |

Table 8.C.2: Steady states, period and amplitude of oscillations for $a = .25$

| $a = 0.5$ | Period (days) | $\overline{E}$ | $\overline{L}$ | Larvae amplitude ($\%\overline{L}$) |
|---|---|---|---|---|
| $b = 2.96$ | 1.72 | 30 | 1.13 | 18.03 |
| $b = 4.22$ | 3.8 | | | 49.8 |
| $b = 8.44$ | 11.5 | | | 109.9 |
| $b = 7.6$ | 10.1 | | | 99.43 |

Table 8.C.3: Steady states, period and amplitude of oscillations for $a = .5$

We provide numerical results for $i \in \{1, 4, 9\}$ and $j \in \{0.1, 0.25, 0.5\}$ initial data close to the steady state (which is drawn in dashed line). Two sets of initial data are chosen, $(E(0), L(0)) = (\overline{E}, \overline{L})$ (green) and $(E(0), L(0)) = (\overline{E}, \overline{L} + 0.02)$ (blue), which gives oscillations that appear to be periodic in time. Simulations are made with $a = a_1$ in Figure 8.C.1 (with $b = b_1^{0.05}$) and a time variable evaluated in $[0, 100]$ days ; $a = a_2$ in Figure 8.C.2 (with $b = b_2^{0.5}$) and a time variable evaluated in $[0, 100]$ days ; $a = a_3$ in Figure 8.C.3 (with $b = b_3^2$) and a time variable evaluated in $[0, 150]$ days.

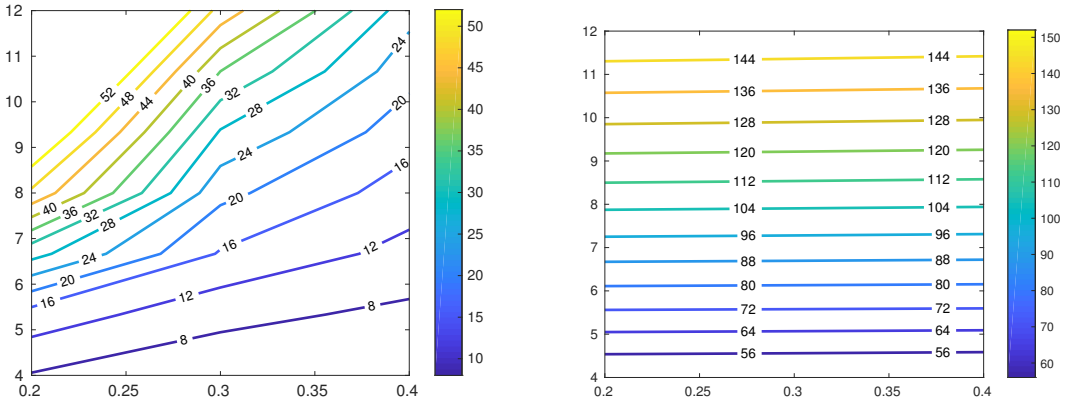


Figure 8.C.4: Larvae dynamics period $T_0$ in days (*left*) and larvae dynamics amplitude (Amp) in percentage of $\overline{L}$ (*right*), for different couples $(a, b)$.

Considering the blue curves, we sum up in the Tables 8.C.1, 8.C.2 and 8.C.3 what we obtain for the period and the oscillations' amplitude taken by the solutions. In the last line of the tables we give a value of $b$ that can be chosen to obtain a period of about 10 days. Relative amplitude of the oscillations is expressed as a percentage of the (constant) value $\overline{L}$.

It is possible to achieve the same period $T_0$ for different couples of parameters $(a, b)$. For a fixed $a$, when $b$ is increasing, the period $T_0$ and the amplitude of larvae are increasing too. The amplitude, on the contrary, mainly depends on $b$. This is illustrated in Figure 8.C.4.

# Chapter 9

# Using sterilizing males to reduce or eliminate *Aedes* populations: insights from a mathematical model

> Les étrangers blêmes, parfois si ridicules, ont beaucoup d'ingéniosité : ils tatouent leurs étoffes de petits signes noirs qui marquent des noms, des rites, des nombres. Et ils peuvent, longtemps ensuite, les rechanter tout à loisir
>
> Victor Segalen, *Les Immémoriaux*.

This chapter is a joint work with Hervé Bossin and Yves Dumont.

**Abstract.** We propose a new mathematical model for population elimination of mosquitoes by means of releases of sterilizing (sterile or incompatible) males, featuring the possibility of an Allee effect. This feature implies that the extinction state is locally stable, and therefore a key object of study appears as the separatrix between the basins of attraction of extinction and wild steady states. We derive exact conditions for population elimination in the case of constant releases, and both sufficient or necessary conditions in the case of impulsive releases. In particular, we obtain analytical estimations of the entrance time into the basin of attraction of the extinction state. The relative importance of the model's parameters is inferred from these results.

Biological parameters are estimated from a case study of an *Aedes polynesiensis* population, for which extensive numerical investigations illustrate the analytical results.

## Introduction

Sterile insect technique (SIT) is a promising technique that has been first studied by E. Knipling and collaborators and first experimented successfully in the early 50's by nearly eradicating screw-worm population in Florida. Since then, SIT has been applied on different pest and disease vectors, like fruit flies or mosquitoes. The classical SIT relies on the mass releases of males sterilized by ionizing radiations. The released sterile males transfer their sterile sperms to wild females, which results in a progressive reduction of the target population. For mosquito control in particular, new approaches stemming from SIT have emerged, namely the RIDL technique, and the *Wolbachia* technique. *Wolbachia* is a bacterium that infects many Arthropods, and among them some mosquito species in nature. It was discovered in 1924 [110]. Since then, particular properties of these bacteria have been unveiled. One of of these properties is particularly useful for vector control: the cytoplasmic incompatibility (CI) property [206, 38]. CI can serve two different control strategies:

- Incompatible Insect Technique (IIT): the sperm of W-males (males infected with CI-inducing *Wolbachia*) is altered so that it can no longer successfully fertilize uninfected eggs. This can result in a progressive reduction of the target population. Thus, when only W-males are released the IIT can be seen as classical SIT. This also supposes that releases are made

regularly until extinction is achieved (when possible) or until a certain threshold is reached (in order to reduce exposure to mosquito bites and the epidemiological risk).

- Population replacement: when males and W-females are released in a susceptible (uninfected) population, and because *Wolbachia* is maternally inherited, W-females will produce offspring while uninfected females won't. This will result in a population replacement by *Wolbachia* infected mosquitoes. It has been showed that this technique may be very interesting with *Aedes aegypti*, shortening their lifespan (see for instance [203]), or more importantly, practically losing their competence for dengue virus transmission [172]. However, it is also acknowledged that *Wolbachia* infection can have drawbacks, like fecundity reduction, so that the use of Wolbachia in the the field can fail [203].

Based on all these biological properties, classical SIT and IIT have been modeled and studied theoretically in a large number of papers in order to derive results to explain the success or not of these strategies using discrete, continuous or hybrid modeling approaches, temporal and spatio-temporal models (see for instance [120, 149, 72, 73, 212, 81, 176, 82] and references therein). More recently, the theory of monotone dynamical systems has been applied efficiently to study SIT and W-SIT systems [35, 197, 9].

The outline of the paper is as follows. First, we explain in Section 9.1 the biological situation considered and the practical questions we want to answer, namely: how to quantify the release effort required to eliminate an *Aedes* population using SIT/IIT, with particular emphasis on the timing and size of the releases. We also justify our modeling choices and give value intervals for most biological parameters in Table 9.1. Then, we perform the theoretical analysis of a simple, compartimentalized population model featuring an Allee effect and a constant sterilizing male population in Section 9.2. Proposition 9.2 gives the bistable asymptotic behavior of the system, and introduces the crucial separatrix between extinction and survival of the population. We also provide analytical inequalities on the entrance time of a trajectory into the extinction set (Proposition 9.3), which is extremely useful to understand what parameters are really relevant and how they interact. We then analyze the model as a control system, after adding a release term. Finally, Section 9.3 exposes numerical investigations of the various models.

In general, all mathematical results are immediately interpreted biologically. To keep the exposition as readable as possible, we gather all technical developments of the proofs into Appendices.

## 9.1 Modeling and biological parameter estimation

### 9.1.1 Modeling context

Our modeling effort is oriented towards an understanding of large-scale time dynamics of a mosquito population in the *Aedes* genus exposed to artificial releases of *sterilizing males*. These males can be either sterilized by irradiation (Sterile Insect Technique approach) or simply have a sterile crossing with wild females due, for instance, to incompatible strains of *Wolbachia* bacteria (Incompatible Insect Technique approach). In either techniques (SIT or IIT), the released males are effectively *sterilizing* the wild females they mate with.

Eggs from mosquitoes of various species in the *Aedes* genus resist to dessication and can wait for months before hatching. Due to rainfall-dependency of natural breeding sites availability, this feature allows for maintaining a large quiescent egg stock through the dry season, which triggers a boom in mosquito abundance when the rainy season resumes. For the populations we model here, natural breeding sites are considered to be prominent, and therefore it is absolutely necessary that our models take the egg stock into account.

We use a system biology approach to model population dynamics. In the present work we neglect the seasonal variations and assume all biological parameters to be constant over time.

Our first compartmental model features egg, larva, adult male and adult female (fertile or sterile) populations. Most transitions between compartments are assumed to be linear. Only three non-linear effects are accounted for.

First, the population size is bounded due to an environmental carrying capacity for eggs, which we model by a logistic term. Secondly, the sterilizing effect creates two sub-populations among inseminated females. Some are inseminated by wild males and become fertile while the others are inseminated by sterilizing males and become sterile. Hence the relative abundance (or more precisely the relative mating power) of sterilizing males with respect to wild males must appear in

the model, and is naturally a nonlinear ratio. Many other parameters may interfere with the mating process for *Aedes* mosquitoes, but this process is not currently totally understood in particular from the male point of view [145, 180], and we stick here to the simplest possible modeling. Thirdly, as a result of sterilizing matings, we expect that the male population can drop down to a very low level. We introduce an Allee effect which come into play in this near-elimination regime. This effect reduces the insemination rate at low male density, as a consequence of difficult mate-finding. It can also be interpreted as a quantification of the size of the mating area relative to the total size of the domain, and compensates in some ways the intrinsic limitations of a mean-field model for a small and dispersed population (cf. [74] and see Remark 9.1). Indeed, we model here temporal dynamics by neglecting spatial variations and assuming homogeneous spatial distribution of the populations. In nature, the distribution of *Aedes* mosquitoes is mostly heterogenous, depending on environmental factors such as vegetation coverage, availability of breeding containers and blood hosts. The proposed simplified homogenous model will thus be exposed to potential criticism.

### 9.1.2 Models and their basic properties

We denote by $E$ the eggs, $L$ the larvae, $M$ the fertile males, $F$ the fertile females and $F_{st}$ the sterile females (either inseminated by sterilizing males or not inseminated at all, due to male scarcity). The time-varying sterilizing male population is denoted $M_i$. We use Greek letters $\mu$ for mortality rates, $\nu$ for transition rates and denote fecundity by $b$ (viable eggs laid per female and per unit of time) and egg carrying capacity by $K$. The full model reads:

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\widetilde{\nu}_E + \mu_E)E, \\ \dfrac{dL}{dt} = \widetilde{\nu}_E E - (\nu_L + \mu_L)L, \\ \dfrac{dM}{dt} = (1-r)\nu_L L - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_L L(1-e^{-\beta(M+\gamma_i M_i)})\dfrac{M}{M+\gamma_i M_i} - \mu_F F, \\ \dfrac{dF_{st}}{dt} = r\nu_L L\big(e^{-\beta(M+\gamma_i M_i)} + \dfrac{\gamma_i M_i}{M+\gamma_i M_i}(1-e^{-\beta(M+\gamma_i M_i)})\big) - \mu_F F_{st}. \end{cases} \tag{9.1}$$

Dynamics of the full system (9.1) is not different from that of the following simplified, three-populations system. We only keep egg, fertile and sterilizing male, and fertile female populations. The value of the hatching parameter $\nu_E$ must be updated to take into account survivorship and development time in the larval stage.

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_E E(1-e^{-\beta(M+\gamma_i M_i)})\dfrac{M}{M+\gamma_i M_i} - \mu_F F. \end{cases} \tag{9.2}$$

The following straightforward lemma means that (9.1) and (9.2) are well-suited for population dynamics modeling since all populations, in these systems, remain positive and bounded.

**Lemma 9.1.** *Let $M_i$ be a non-negative, piecewise continuous function on $\mathbb{R}_+$. The solution to the Cauchy problems associated with (9.1), (9.2) and non-negative initial data is unique, exists on $\mathbb{R}_+$, is continuous and piecewise continuously differentiable. This solution is also forward-bounded and remains non-negative. It is positive for all positive times if $F(0)$ or $E(0)$ (or also $L(0)$ in the case of (9.1)) is positive.*

In addition, these systems are monotone in the sense of the monotone systems theory (see [208]).

**Lemma 9.2.** *The system (9.2) is monotone on the set $\mathcal{E}_3 := \{E \leq K\} \subset \mathbb{R}^3_+$ for the order induced by $\mathbb{R}^3_+$ and the restriction of system (9.1) to the four first coordinates (omitting $F_{st}$, which does*

*not appear in any other compartment) is monotone on the set* $\mathcal{E}_4 := \{E \leq K\} \subset \mathbb{R}^4_+$ *for the order induced by* $\mathbb{R}^4_+$.

*Moreover,* $\mathcal{E}_3$ *(respectively* $\mathcal{E}_4$*) is forward invariant for (9.2) (respectively for the restriction of (9.1) to the four first coordinates), and any trajectory enters it in finite time.*

*Proof.* We compute the Jacobian matrix of the system (9.2):

$$J = \begin{pmatrix} -\frac{bF}{K} - (\nu_E + \mu_E) & 0 & b(1 - \frac{E}{K}) \\ (1-r)\nu_E & -\mu_M & 0 \\ r\nu_E(1 - e^{-\beta(M+\gamma_i M_i)})\frac{M}{M+\gamma_i M_i} & \frac{r\nu_E E}{M+\gamma_i M_i}(\beta M e^{-\beta(M+\gamma_i M_i)} + (1 - e^{-\beta(M+\gamma_i M_i)})\frac{\gamma_i M_i}{M+\gamma_i M_i}) & -\mu_F \end{pmatrix}.$$

It has non-negative extra-diagonal coefficients on $\mathcal{E}_3$, which proves that the system is indeed monotone on this set. In addition, if $E(t_0) > K$ then let $T[t_0] := \{t \geq t_0, \forall t' \in [t_0, t), E(t') > K\} \subset \mathbb{R}$. Let $T^+[t_0] := \sup T[t_0]$. For any $t \in T[t_0]$ we have $\dot{E}(t) \leq -(\nu_E + \mu_E)E(t)$. Hence by integration we find that $T^+[t_0] \leq t_0 + \frac{1}{\nu_E + \mu_E} \log(K/E(t_0)) < +\infty$, which proves Lemma 9.2 (the proof being similar for the claims on (9.1)). □

**Remark 9.1.** *The Allee effect term* $1 - \exp(-\beta M)$ *can also be interpreted in the light of [74]. This is the probability that an emerging female finds a male to mate with in her neighborhood.*

*Using a "mean-field" model of ordinary differential equations here is certainly debatable, since in the case of population extinction the individuals may eventually be very dispersed, and heterogeneity would play a very important role. However, we think that getting a neat mathematical understanding of the simplest system we study here is a necessary first step before moving to more complex systems. The Allee term compensates, as far as the qualitative behavior is concerned, what the model structurally lacks. Here, we are able to perform proofs and analytical computations. This gives a starting point for benchmarking what to expect as an output of release programs using sterilizing males, according to the models.*

### 9.1.3 Parameter estimation from experimental data

| Symbol | Name | Value interval | Source |
|---|---|---|---|
| $r_{\text{viable}}$ | Proportion of viable eggs | $95 - 99\%$ | Field collection, [105, p. 121] |
| $N_{\text{eggs}}$ | Number of eggs laid per laying | $55 - 75$ | [195] |
| $\tau_{\text{gono}}$ | Duration of gonotrophic cycle | $4 - 7$ days | [124, 213, 195] |
| $\tau_E$ | Egg half-life | $15 - 30$ days | Estimation (to be determined) |
| $\tau_L$ | Time from hatching to emergence | $8 - 11$ days | Lab data, [105, p. 104] |
| $r_L$ | Survivorship from larva first instar to pupa | $67 - 69\%$ | Lab data, [105, p. 106] |
| $r$ | Sex ratio (male:female) | $49\%$ | Production data (ILM) |
| $\tau_M$ | Adult male half-life | $5 - 9$ days | Lab data, [105, p. 50] |
| $\tau_F$ | Adult female half-life | $15 - 21$ days | Lab data, [105, p. 50] |
| $\gamma_i$ | Mating competitiveness of sterilizing males | 1 | Lab [105, pp. 51–53], field [179] |

Table 9.1: Parameter values for some populations of *Aedes polynesiensis* in French Polynesia at a temperature of $27°C$.

For numerical simulations, we use experimental (lab and field) values of the biological parameters in (9.1)-(9.2). We consider specifically a population of *Aedes polynesiensis* in French Polynesia which has been studied in [124, 213, 195], and more recently in [49, 106, 107, 105].

Values of most parameters are given in Table 9.2, and are deduced from experimental data gathered in Table 9.1. Some data come from unpublished results obtained at Institut Louis Malardé during the rearing of *Aedes polynesiensis* for a pilot IIT program. They are labelled as "Production data (ILM)". Note that we do not give values for $\beta$ and $\widetilde{\nu}_E$ because they are very hard to estimate. Ongoing experiments of one of the author may help approximating them in the future for this *Aedes polynesiensis* population. Finally when it exists, we use the knowledge about population size (male and female) granted by mark-release-recapture experiments to adjust the environmental carrying capacity $K$ for population and season.

| Symbol | Name | Formula | Value interval |
|---|---|---|---|
| $b$ | Effective fecundity | $\frac{r_{\text{viable}} N_{\text{eggs}}}{\tau_{\text{gono}}}$ | $7.46 - 14.85$ |
| $\mu_L$ | Larva death rate | $\frac{-\log(r_L)}{\tau_L}$ | $0.034 - 0.05$ |
| $\nu_L$ | Larva to adult transition rate | $\frac{1}{\tau_L}$ | $0.09 - 0.125$ |
| $\frac{\nu_E}{\widetilde{\nu}_E}$ | Larval coefficient for effective hatching rate | $\frac{\nu_L}{\nu_L + \mu_L}$ | $0.64 - 0.79$ |
| $\mu_E$ | Egg death rate | $\frac{\log(2)}{\tau_E}$ | $0.023 - 0.046$ |
| $\mu_M$ | Adult male death rate | $\frac{\log(2)}{\tau_M}$ | $0.077 - 0.139$ |
| $\mu_F$ | Adult female death rate | $\frac{\log(2)}{\tau_F}$ | $0.033 - 0.046$ |

Table 9.2: Conversion of the biological parameter from Table 9.1 into mathematical parameters for systems (9.1) and (9.2)

## 9.2 Theoretical study of the simplified model

For later use, we introduce the usual relations $\ll$, $<$ and $\leq$ on $\mathbb{R}^d$ (where $d \geq 1$) as the coordinate-wise partial orders on $\mathbb{R}^d$ induced by the cone $\mathbb{R}^d_+$. More precisely, for $x, y \in \mathbb{R}^d$,

- $x \leq y$ if and only if for all $1 \leq i \leq d$, $x_i \leq y_i$,
- $x < y$ if and only if $x \leq y$ and $x \neq y$,
- $x \ll y$ if and only if for all $1 \leq i \leq d$, $x_i < y_i$.

### 9.2.1 Constant incompatible male density

First we study system (9.2) with constant incompatible male density $M_i(t) \equiv M_i$.

We introduce the three scalars

$$\mathcal{N} := \frac{br\nu_E}{\mu_F(\nu_E + \mu_E)}, \quad \lambda := \frac{\mu_M}{(1-r)\nu_E K}, \quad \psi := \frac{\lambda}{\beta} \tag{9.3}$$

and define the function $f : \mathbb{R}^2_+ \to \mathbb{R}$, with the two parameters $\mathcal{N}$ and $\psi$:

$$f(x, y) := x(1 - \psi x)(1 - e^{-(x+y)}) - \frac{1}{\mathcal{N}}(x + y). \tag{9.4}$$

The two aggregated numbers, $\mathcal{N}$ and $\psi$ essentially contain all the information about system (9.2): $\mathcal{N}$ is the classical basic offspring number, $\psi$ is the ratio between the typical male population size at which the Allee effect comes into play and the male population size at wild equilibrium, as prescribed by the egg carrying capacity.

The ODE system (9.2) has simple dynamical properties because it is monotone and we can count its steady states and even know their local stability. Let $M_i \geq 0$. It is straightforward to show that system (9.2) always admits a trivial steady-state $(0, 0, 0)$ and eventually one (at least) non-trivial steady state $(E^*, M^*, F^*) \in \mathbb{R}^3_+$ solution of

$$E = \frac{b}{\nu_E + \mu_E} F(1 - \frac{E}{K}), \quad E = \frac{\mu_M}{(1-r)\nu_E} M, \quad F = \frac{r\nu_E}{\mu_F} E(1 - e^{-\beta(M + \gamma_i M_i)}) \frac{M}{M + \gamma_i M_i}.$$

Using the first two equation into the third one yields

$$\frac{\mu_F(\nu_E + \mu_E)}{br\nu_E}(M + \gamma_i M_i) = M(1 - \frac{\mu_M}{(1-r)\nu_E K} M)(1 - e^{-\beta(M+\gamma_i M_i)}),$$

from which we deduce

$$\begin{cases} E^* = K\lambda M^*, \\ F^* = \dfrac{K(\nu_E + \mu_E)}{b} \dfrac{\lambda M^*}{1 - \lambda M^*}, \\ f(\beta M^*, \gamma_i \beta M_i) = 0. \end{cases}$$

Hence for a given value $M_i \geq 0$, the number of steady states of (9.2) is equal to the number of positive solutions $M^*$ to $f(\beta M^*, \beta\gamma_i M_i) = 0$, plus 1. The trivial steady state $(0,0,0)$ is also locally asymptotically stable (LAS). The following lemma give us additional informations about the positive steady state(s):

**Lemma 9.3.** *Assume $\mathcal{N} > 4\psi$. Let $\theta_0 \in (0,1)$ be the unique solution to $1-\theta_0 = -\frac{4\psi}{\mathcal{N}}\log(\theta_0)$, and*

$$M_i^{crit} := \frac{1}{\gamma_i \beta} \max_{\theta\in[\theta_0,1]} \Big( -\log(\theta) - \frac{1}{2\psi}\big(1 - \sqrt{1 + \frac{4\psi}{\mathcal{N}}\frac{\log(\theta)}{1-\theta}}\big)\Big).$$

*If $M_i^{crit} > 0$ then (9.2) has:*

- *0 positive steady state if $M_i > M_i^{crit}$,*
- *2 positive steady states $\mathbf{E}_- \ll \mathbf{E}_+$ if $M_i \in [0, M_i^{crit})$,*
- *1 positive steady state $\mathbf{E}$ if $M_i = M_i^{crit}$.*

*In addition, $\mathbf{E}_-$ is unstable and $\mathbf{E}_+$ is locally asymptotically stable. If $M_i^{crit} < 0$ then (9.2) has no positive steady state, and if $M_i^{crit} = 0$ then there exists a unique positive steady state. In particular, if $\mathcal{N} \leq 1$ then $M_i^{crit} < 0$.*

*On the contrary, if $\mathcal{N} \leq 4\psi$ then there is no positive steady state.*

*Proof.* Let us give a quick overview of the remainder of the proof, which is detailed in Appendix 9.A, page 179. We are going to study in details the solutions $(x,y)$ to $f(x,y) = 0$. First, we prove that $x < 1/\psi$. Then, we check that for any $y > 0$, $x \mapsto f(x,y)$ is either concave or convex-concave. In addition, it is straightforward that $f(0,y) < 0$ and $\lim_{x\to+\infty} f(x,y) = -\infty$, so that for any $y > 0$, we conclude that there are either 0, 1 or 2 real numbers $x > 0$ such that $f(x,y) = 0$.

Then, we introduce $\xi = 4\psi/\mathcal{N}$. In fact, in order to determine $(x,y) \in \mathbb{R}_+^2$ such that $f(x,y) = 0$ we can introduce $\theta = e^{-(x+y)}$ and then check easily that $y = h_\pm(\theta)$, where

$$h_\pm(\theta) = -\log(\theta) - \frac{1}{2\psi} \pm \frac{1}{2\psi}\sqrt{1 + \xi\frac{\log(\theta)}{1-\theta}}. \tag{9.5}$$

Let $\theta_0(\xi)$ be the unique solution in $(0,1)$ to $1-\theta_0(\xi) = -\xi\log(\theta_0(\xi))$, and

$$\alpha^{\mathrm{crit}}(\xi,\mathcal{N}) := \max_{\theta\in[\theta_0(\xi),1]} -\log(\theta) - \frac{1}{2\psi}\big(1 - \sqrt{1 + \xi\frac{\log(\theta)}{1-\theta}}\big). \tag{9.6}$$

Collecting the previous facts, and studying the function $h_\pm$ (see Appendix 9.A.2, page 180), we can prove that the next point of Lemma 9.3 holds with the threshold $M_i^{\mathrm{crit}} = \frac{\mathcal{N}}{4\psi\beta\gamma_i}\alpha^{\mathrm{crit}}(\xi,\mathcal{N})$.

We remark that if $\mathcal{N} \leq 1$ then it is easily checked that $M_i^{\mathrm{crit}} < 0$, using the fact that if $\alpha \in (0,1)$ then $\sqrt{1-\alpha} \leq (1-\alpha)/2$. If $\theta \in (\theta_0, 1)$ then $\frac{4\psi\log(\theta)}{\mathcal{N}(1-\theta)} < 1$, and therefore

$$-\log(\theta) - \frac{1}{4\psi}\big(1 - \sqrt{1 + \frac{4\psi}{\mathcal{N}}\frac{\log(\theta)}{1-\theta}}\big) \leq -\log(\theta)\big(1 - \frac{1}{\mathcal{N}}\big) - \frac{1}{4\psi} < 0.$$

In the final part of the proof, we show that 0 is always locally stable and then treat separately the cases $M_i = 0$ and $M_i > 0$, showing that, when they exist, the greater positive steady state is locally stable while the smaller one is unstable.

□

**Remark 9.2.** *In Lemma 9.3, the condition to have at least one positive equilibrium, $\mathcal{N} > 4\psi$, is very interesting and particularly makes sense when rewritten as $\frac{\mathcal{N}}{\lambda} > \frac{4}{\beta}$. Indeed $\frac{\mathcal{N}}{\lambda}$ can be seen as the theoretical male progeny at next generation, starting from wild equilibrium. If this amount is large enough (larger than some constant times the population size at which the Allee effect comes into play) then the population can maintain. In any case, if this condition is not satisfied, then the population collapses. For the population to maintain: either the fitness is good and thus $\mathcal{N}$ is very large, or the probability of one female to mate is high and thus $1/\beta$ is small. However, whatever the values taken by $\mathcal{N}$ and $\beta$, if, for any reason, the male population at equilibrium decays, the population can be controlled and possibly collapses.*

**Remark 9.3.** *If $\beta$ is not too small, then the "wild" steady state is approximately given by $M^*(M_i = 0) \simeq \frac{1}{\lambda}(1 - \frac{1}{\mathcal{N}})$ and the critical sterilizing level is approximately $M_i^{crit} \simeq \widetilde{y} = \frac{\mathcal{N}}{4\lambda\gamma_i}\left(1 - \frac{1}{\mathcal{N}}\right)^2$ (see the definition in Appendix 9.A, in particular we know that $M_i^{crit} \leq \widetilde{y}$). As a consequence, the target minimal constant density of sterilizing males compared to wild males in order to get unconditional extinction (i.e. to make $(0,0,0)$ globally asymptotically stable, see Proposition 9.1, page 167) is well approximated by the simple formula*

$$\rho^* := \frac{M_i^{crit}}{M^*(M_i = 0)} \simeq \frac{\mathcal{N} - 1}{4\gamma_i}.$$

*With the values from Tables 9.1 and 9.2, for $\gamma_i = 1$ (this means that introduced male are as competitive as wild ones for mating with wild females), we find*

$$\rho^* \in \left(\frac{7.46 \cdot 0.46 \cdot \nu_E}{4 \cdot 0.046 \cdot (\nu_E + 0.046)} - 0.25, \frac{14.85 \cdot 0.48 \cdot \nu_E}{4 \cdot 0.033 \cdot (\nu_E + 0.023)} - 0.25\right)$$

*For instance, if $\nu_E = 0.01$ then this interval is $(3.5, 22, 7)$, if $\nu_E = 0.05$ then this interval is $(10.6, 51.7)$ and if $\nu_E = 0.1$ then this interval is $(14.1, 61.4)$. As $\nu_E$ goes to $+\infty$, the interval goes to $(20.7, 75.7)$. This example agrees with standard SIT Protocol that indicates to release at least 10 times more sterile males than wild males, recalling that here we deal with a highly reproductive species (with the above values, the lowest estimated basic reproduction number is 14.9, obtained for $\nu_E = 0.01$).*

Asymptotic dynamics are easily deduced from the characterization of steady states and local behavior of the system (Lemma 9.3), because of the monotonicity (see [208]).

**Proposition 9.1.** *If (9.2) has only the steady state $(0,0,0)$ then it is globally asymptotically stable.*

*If there are two other steady states $\mathbf{E}_- \ll \mathbf{E}_+$ then almost every orbit converges to $\mathbf{E}_+$ or $(0,0,0)$. Let $K_+ := [(0,0,0), \mathbf{E}_+]$. The compact set $K_+$ is globally attractive and positively invariant. The basin of attraction of $(0,0,0)$ contains $[0, \mathbf{E}_-)$ and the basin of attraction of $\mathbf{E}_+$ contains $(\mathbf{E}_-, \infty)$.*

Now that we have established that the system is typically bistable, the main object to investigate is the separatrix between the two basins of attraction. This is the aim of the next proposition.

**Proposition 9.2.** *Assume $M_i^{crit} > 0$ and $M_i \in [0, M_i^{crit})$.*

*Then there exists a separatrix $\Sigma \subset \mathbb{R}^3_+$, which is a sub-manifold of dimension 2, such that for all $X \neq Y \in \Sigma$, $X \not\leq Y$ and $Y \not\leq X$, and for all $\widehat{X} \in \Sigma$, $X_0 > \widehat{X}$ implies that $X(t)$ converges to $\mathbf{E}_+$, and $X_0 < \widehat{X}$ implies that $X(t)$ converges to $\mathbf{0}$. In particular, $\mathbf{E}_- \in \Sigma$.*

*Let $\Sigma_+ := \left\{X \in \mathbb{R}^3_+, \exists \widehat{X} \in \Sigma, X > \widehat{X}\right\}$ and $\Sigma_- := \left\{X \in \mathbb{R}^3_+, \exists \widehat{X} \in \Sigma, X < \widehat{X}\right\}$. Then $\mathbb{R}^3_+ = \Sigma_- \cup \Sigma \cup \Sigma_+$, $\Sigma_+$ is the basin of attraction of $\mathbf{E}_+$ and $\Sigma_-$ is the basin of attraction of $\mathbf{0}$.*

*In addition, there exists $E_M, F_M > 0$ such that*

$$\Sigma_- \subset \left\{X \in \mathbb{R}^3_+, \quad X_1 \leq E_M, X_3 \leq F_M\right\}.$$

**Remark 9.4.** *In order to reach extinction, the last point of Proposition 9.2 states that both egg and fertile female populations must stand simultaneously below given thresholds. This obvious fact receives here a mathematical quantification. With simple words: no matter how low the fertile female population $F$ has dropped, if there remains at least $E_M$ eggs then the wild population will recover.*

*Proposition 9.2.* We state a preliminary fact: For all $v^0 \in \{v \in \mathbb{R}^3_+, \forall i, v_i > 0, \sum_i v_i = 1\} =: \mathcal{S}^2_+$, there exists a unique $\rho_0(v^0)$ such that the solution to (9.2) with initial data $\rho v^0$ converges to $\mathbf{0}$ if $\rho < \rho_0(v^0)$ and to $\mathbf{E}_+$ if $\rho > \rho_0(v^0)$.

This fact comes from the strict monotonicity of the system, and from the estimate $\rho_0(v^0) \leq \max_i \frac{v_i^0}{(\mathbf{E}_-)_i} < +\infty$, combined with Proposition 9.1.

Then we claim that $\Sigma = \{\rho_0(v^0)v^0, \quad v^0 \in \mathcal{S}^2_+\}$. The direct inclusion is a corollary of the previous fact. The converse follows from the fact that $\Sigma_\pm$, being the basins of attraction of attracting points, are open sets.

The remainder of the proof consists of a simple computation showing that if $F_0$ or $E_0$ is large enough then for some $t > 0$ we have $(E, M, F)(t) > \mathbf{E}_-$. In details, we can prove that if $F_0$ is large enough then for any $E_0, M_0$ and $\epsilon > 0$, we can get $E(s) \geq (1-\epsilon)K$ for $s \in (t_0(\epsilon, E_0, F_0), t_1(\epsilon, E_0, F_0))$, where $t_0$ is decreasing in $F_0$ and $t_1$ is increasing in $F_0$ and unbounded as $F_0$ goes to $+\infty$. Then, if $E > (1-\epsilon)K$ for $\epsilon$ small enough on a large enough time-interval, we deduce $M(t) > (1-\epsilon)^2(1-r)\frac{\nu_E}{\mu_M}K$ for some $t > 0$. Upon choosing $\epsilon$ small enough and $F_0$ large enough we finally get $(E, M, F)(t) > \mathbf{E}_-$. The scheme is similar when taking $E_0$ large enough. □

At this stage, we know that starting from the positive equilibrium, and assuming that the population of sterile males $M_i$ is greater than $M_i^{\text{crit}}$, the solution will reach the basin of attraction of the trivial equilibrium in a finite time, $\tau(M_i)$. We obtain now quantitative estimates on the duration of this transitory regime. Rigorously, we define

$$\tau(M_i) := \inf\big\{t \geq 0,\ (E, M, F)(t) \in \Sigma_-(M_i = 0),\\ \text{where } (E, M, F)(0) = \mathbf{E}_+(M_i = 0) \text{ and } (E, M, F) \text{ satisfies (9.2)}\big\}. \quad (9.7)$$

We obtain simple upper and lower bounds for $\tau(M_i)$ in terms of various parameters:

**Proposition 9.3.** *Let $M_i > M_i^{crit}$, and $Z = Z(\psi)$ be the unique real number in $(0, \frac{1}{2\psi})$ such that*

$$e^{-Z} = \frac{\psi}{1+\psi-\psi Z},$$

*and $Z_0 := 1 + \psi - \psi Z$. Then*

$$\tau(M_i) \geq \frac{1}{\mu_F}\log\big(1 + \frac{\mathcal{N}^2(1-\psi Z)^3}{\psi Z Z_0^2} - \frac{\mathcal{N}(1-\psi Z)}{\psi Z Z_0}\big). \quad (9.8)$$

*Let $\sigma = sgn(\nu_E + \mu_E - \mu_F)$, $\sigma_E := \mu_M/(\nu_E + \mu_E)$ and $\sigma_F := \mu_M/\mu_F$. If $\epsilon := \frac{M_+^*}{M_+^* + M_i} < 1/\mathcal{N}$, let*

$$g(\epsilon) := \sqrt{1 + \frac{4\mathcal{N}\sigma_E\sigma_F\epsilon}{(\sigma_F - \sigma_E)^2}}.$$

*Assume that $\sigma_F, \sigma_E > 1$,*

$$g(\epsilon)\sigma(\sigma_F - \sigma_E) < \max\big((2\mathcal{N}-1)\sigma_F + \sigma_E, (2\sigma_E - 1)\sigma_F\big), \quad (\sigma_F - 1)(\sigma_E - 1) > \epsilon\mathcal{N}.$$

*Then*

$$\tau(M_i) \leq \frac{2\sigma_E}{\mu_F\big(\sigma_F + \sigma_E - g(\epsilon)\sigma(\sigma_F - \sigma_E)\big)}\log\Big(\frac{\mathcal{N}-1}{\psi}\Big(\frac{(\mathcal{N}-1)\sigma_F + 1 - \epsilon\mathcal{N}}{(\sigma_F - 1)(\sigma_E - 1) - \epsilon\mathcal{N}}\\ + \frac{\sigma_E\sigma_F\big(g(\epsilon)\sigma(\sigma_F - \sigma_E) + (2\mathcal{N}-1)\sigma_F + \sigma_E\big)}{\big(2\sigma_E\sigma_F - (\sigma_E + \sigma_F) + \sigma(\sigma_F - \sigma_E)g(\epsilon)\big)g(\epsilon)\sigma(\sigma_F - \sigma_E)}\Big)\Big). \quad (9.9)$$

*Proof.* The proof relies on explicit computation of sub- and super-solutions, detailed in Appendix 9.B. □

**Remark 9.5.** *The dependency in $\psi$ of Proposition 9.3's upper estimate on $\tau$ is approximately equal to $\frac{1}{\min(\nu_E + \mu_E, \mu_F)}$. One order of magnitude of $\psi$ (the ratio between the wild population size and the Allee population size) therefore typically corresponds to the maximum of one adult female and one egg lifespan in terms of release duration needed to get extinction.*

**Remark 9.6.** *At this stage, we obtain an analytic upper bound only in the case of massive releases ($\epsilon$ small enough). A more refined upper bound could theoretically be obtained, see the derivation in Appendix 9.B, in particular Lemma 9.12.*

### 9.2.2 Adding a control by means of releases

In a slightly more realistic model, the level of sterilizing male population should vary with time, depending on the releases $t \mapsto u(t) \geq 0$ and on a fixed death rate $\mu_i$. This model reads

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dM_i}{dt} = u(t) - \mu_i M_i, \\ \dfrac{dF}{dt} = r\nu_E E(1-e^{-\beta(M+\gamma_i M_i)})\dfrac{M}{M+\gamma_i M_i} - \mu_F F. \end{cases} \tag{9.10}$$

In (9.10), the number of sterilizing males released between times $t_1$ and $t_2 > t_1$ is simply equal to $\int_{t_1}^{t_2} u(t)dt$.

First, if the release is *constant*, say $u(t) \equiv u_0$, then $M_i(t) = e^{-\mu_i t}M_i^0 + \frac{u_0}{\mu_i}(1-e^{-\mu_i t})$. The special case $M_i^0 = \frac{u_0}{\mu_i}$ leads back to system (9.2), with $M_i \equiv M_i^0$. For general $M_i^0 \geq 0$, we notice that $M_i(t)$ converges to $\frac{u_0}{\mu_i}$ as $t$ goes to $+\infty$.

**Proposition 9.4.** *Assume $u(t) \equiv u_0$.*

*If $u_0 > \mu_i M_i^{crit}$ (defined in Lemma 9.3) then $\mathbf{0}$ is globally asymptotically stable.*

*If $u_0 < \mu_i M_i^{crit}$, then there exists open sets $\Sigma_-(u_0), \Sigma_+(u_0) \subset \mathbb{R}^4_+$, respectively the basins of attraction of $\mathbf{0}$ and $\mathbf{E}_+$ (defined for (9.2) with $M_i = \frac{u_0}{\mu_i}$), separated by a set $\Sigma(u_0)$ which enjoys the same properties as those of $\Sigma(0)$, listed in Proposition 9.2.*

(We do not treat the case $u_0 = \mu_i M_i^{\mathrm{crit}}$).

*Proof.* Since system (9.10) is monotone with respect to the control $u$ (with sign pattern $(-,-,-,+)$), we can use Lemma 9.3 and Proposition 9.2 with sub- and super-solution to get this result in a straightforward way. □

From now on we will restrict ourselves to (possibly truncated) time-periodic controls, which means that we assume that there exists $N_r \in \mathbb{Z}_+ \cup \{+\infty\}$ (the number of release periods), a period $T > 0$ and a function $u_0 : [0,T] \to \mathbb{R}_+$ such that

$$u(t) = \begin{cases} u_0(t-nT) & \text{if } nT \leq t < (n+1)T \text{ for some } N_r > n \in \mathbb{Z}_+, \\ 0 & \text{otherwise.} \end{cases} \tag{9.11}$$

We use the notation $u \equiv [T, u_0, N_r]$ to describe this control $u$.

As before, we can compute in case (9.11)

$$\begin{aligned} M_i(t) &= e^{-\mu_i t}M_i^0 + \int_0^t u(t')e^{-\mu_i(t-t')}dt' \\ &= e^{-\mu_i t}\Big(M_i^0 + \frac{e^{\mu_i(\lfloor \frac{t}{T} \rfloor \wedge N_r)T}-1}{e^{\mu_i T}-1}\int_0^T u_0(t')e^{\mu_i t'}dt' + \int_{T(\lfloor \frac{t}{T} \rfloor \wedge N_r)}^t u(t')e^{\mu_i t'}dt'\Big) \end{aligned}$$

(Here, for $a, b \in \mathbb{Z}$, we let $a \wedge b = \min(a,b)$).

If $N_r = +\infty$, for any $u_0 \neq 0$ there exists a unique periodic solution $M_i$, uniquely defined by its initial value

$$M_i^{0,\mathrm{per}} = \frac{1}{1-e^{-\mu_i T}}\int_0^T u_0(t')e^{\mu_i t'}dt',$$

and which we denote by $M_i^{\mathrm{per}}[u_0]$.

**Lemma 9.4.** *Solutions to (9.10) with $u \equiv [T, u_0, +\infty]$ are such that $M_i$ converges to $M_i^{per}[u_0]$, and the other compartments converge to a solution of*

$$\begin{cases} \dfrac{dE}{dt} = bF(1 - \dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_E E(1 - e^{-\beta(M+\gamma_i M_i^{per}[u_0])})\dfrac{M}{M + \gamma_i M_i^{per}[u_0]} - \mu_F F. \end{cases} \tag{9.12}$$

*Convergence takes place in the sense that the $L^\infty$ norm on $(t, +\infty)$ of the difference converges to 0 as $t$ goes to $+\infty$.*

*Proof.* Convergence of $M_i$ is direct from the previous formula. Then, as for Proposition 9.4 the monotonicity of the system implies the convergence. □

Let $\overline{M}_i[u_0] := \max M_i^{\text{per}}[u_0]$ and $\underline{M}_i[u_0] := \min M_i^{\text{per}}[u_0]$.

**Proposition 9.5.** *If $\underline{M}_i[u_0] > M_i^{crit}$ then $\mathbf{0}$ is globally asymptotically stable for (9.12).*

*On the contrary, if $\overline{M}_i[u_0] < M_i^{crit}$ then (9.12) has at least one positive periodic orbit. In this case the basin of attraction of $\mathbf{0}$ contains the interval $\big(\mathbf{0}, \mathbf{E}_-(M_i = \underline{M}_i[u_0])\big)$, and any initial data above $\mathbf{E}_+(M_i = \underline{M}_i[u_0])$ converges to $\overline{X}^{per}[u_0]$.*

*Proof.* System (9.12) is a periodic monotone dynamical system. It admits a unique non-negative solution $X = (E, M, F)$. In fact, we consider the constant sterile population model

$$\begin{cases} \dfrac{dE_m}{dt} = bF_m\left(1 - \dfrac{E_m}{K}\right) - (\nu_E + \mu_E)E_m, \\ \dfrac{dM_m}{dt} = (1-r)\nu_E E_m - \mu_M M_m, \\ \dfrac{dF_m}{dt} = r\nu_E \dfrac{M_m}{M_m + \underline{M}_i[u_0]}\big(1 - e^{-\beta(M_m + \gamma_i \underline{M}_i[u_0])}\big)E_m - \mu_F F_m. \end{cases} \tag{9.13}$$

such that, using a comparison principle, the solution $X_m = (E_m, M_m, F_m)$ verifies $X_m \geq X$ for all time $t > 0$. Thus if $X_m$ converges to $\mathbf{0}$, so will $X$. The behavior of system (9.13) follows from the results obtained in the previous section. A sufficient condition to have $\mathbf{0}$ globally asymptotically stable in (9.12) is therefore given by $\underline{M}_i^{\text{per}} > M_i^{\text{crit}}$.

The remainder of the claim is better seen at the level of the discrete dynamical system defined by (9.12). Periodic orbits are in one-to-one correspondence with the fixed points of the monotone mapping $\Phi[u_0] : \mathbb{R}_+^3 \to \mathbb{R}_+^3$ defined as the Poincaré application of (9.12) (mapping an initial data to the solution at time $t = T$). Now, if $X^* := (E^*, M^*, F^*)$ is the biggest (*i.e.* stable) steady state of (9.2) at level $M_i = \overline{M}_i[u_0] < M_i^{\text{crit}}$, then for any $(E, M, F) \gg (E^*, M^*, F^*)$ and $M_i' \leq M_i$, writing the right-hand side as $\Psi = (\Psi_1, \Psi_2, \Psi_3)$ we have

$$\Psi_1(E^*, M, F, M_i') > 0,$$

$$\Psi_2(E, M^*, F, M_i') > 0,$$

$$\Psi_3(E, M, F^*, M_i') > 0.$$

In other words, the interval $\big(X^*, +\infty\big)$ is a positively invariant set. Therefore, $\Phi[u_0](X^*) > X^*$. Thus the sequence $\big(\Phi[u_0]^k(X^*)\big)_k$ is increasing and bounded in $\mathbb{R}_+^3$: it must converge to some $\underline{X}^* > X^*$. The same reasoning (with reversed inequalities) applies with the sequence starting at the stable equilibrium associated with $M_i = \underline{M}_i[u_0]$: it must decrease, and thus converge to some $\overline{X}^* \geq \underline{X}^*$.

By our proof we have shown that the open interval $\big(\mathbf{E}_+(M_i = \underline{M}_i[u_0]), +\infty\big)$ belongs to the basin of attraction of $\overline{X}^{\text{per}}$, and we can also assert that $\big(\mathbf{E}_-(M_i = \overline{M}_i[u_0]), \mathbf{E}_+(M_i = \overline{M}_i[u_0])\big)$ belongs to the basin of attraction of $\underline{X}^{\text{per}}$, while as usual $\big(\mathbf{0}, \mathbf{E}_-(M_i = \underline{M}_i[u_0])\big)$ is in the basin of attraction of $\mathbf{0}$. □

By a direct application of the previous results

**Lemma 9.5.** *If $\underline{M}_i[u_0] > M_i^{crit}$ then the control $u \equiv [T, u_0, n]$ (with $n \in \mathbb{Z}_+$) leads to extinction (i.e. the solution with initial data $\mathbf{E}_+$ goes to 0 as $t$ goes to $+\infty$) as soon as*

$$n \geq \frac{\tau(\underline{M}_i[u_0])}{T}. \tag{9.14}$$

A special case of (9.10)-(9.12) is obtained by choosing $u_0 = u_0^\epsilon = \frac{\Lambda}{\epsilon}\mathbb{1}_{[0,\epsilon]}$ for some $\Lambda > 0$ and letting $\epsilon$ go to 0. Then there exists a unique limit as $\epsilon$ goes to 0, which is given by the following impulsive differential system derived from (9.10):

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dM_i}{dt} = -\mu_i M_i, \\ M_i(nT^+) = M_i(nT) + \Lambda \text{ for } n \in \mathbb{Z}_+ \text{ with } 0 \leq n < N_r, \\ \dfrac{dF}{dt} = r\nu_E E(1 - e^{-\beta(M+\gamma_i M_i)})\dfrac{M}{M+\gamma_i M_i} - \mu_F F. \end{cases} \tag{9.15}$$

In (9.15), $M_i$ converges to the periodic solution

$$M_i^{\text{imp}}(t) := \lim_{\epsilon \to 0} M_i^{\text{per}}[u_0^\epsilon] = \frac{\Lambda e^{-\mu_i(t - \lfloor \frac{t}{T} \rfloor T)}}{1 - e^{-\mu_i T}}$$

We can compute explicitly $\underline{M}_i^{\text{imp}} := \frac{\Lambda e^{-\mu_i T}}{1-e^{-\mu_i T}}$ and $\overline{M}_i^{\text{imp}} := \frac{\Lambda}{1-e^{-\mu_i T}}$, respectively the minimum and the maximum of $M_i^{\text{imp}}$. We also define the following periodic monotone system as a special case of (9.12):

$$\begin{cases} \dfrac{dE}{dt} = bF(1-\dfrac{E}{K}) - (\nu_E + \mu_E)E, \\ \dfrac{dM}{dt} = (1-r)\nu_E E - \mu_M M, \\ \dfrac{dF}{dt} = r\nu_E E(1 - e^{-\beta(M+\gamma_i M_i^{\text{imp}})})\dfrac{M}{M+\gamma_i M_i^{\text{imp}}} - \mu_F F. \end{cases} \tag{9.16}$$

The right-hand side of system (9.15) is locally Lipschitz continuous on $\mathbb{R}^3$. Thus, using a classic existence theorem (Theorem 1.1, p. 3 in [21] ), there exists $T_e > 0$ and a unique solution defined from $(0, T_e) \to \mathbb{R}^3$. Using standard arguments, it is straightforward to show that the positive orthant $\mathbb{R}^3_+$ is an invariant region for system (9.15).

We estimate the (minimum) size of the releases $\Lambda$ and periodicity $T$, such that the wild population goes to extinction.

**Proposition 9.6.** *Let $\mathcal{S} := \dfrac{(1-r)\nu_E \mathcal{N}}{4\mu_M \gamma_i}\left(1 - \dfrac{1}{\mathcal{N}}\right)^2 K$. If*

$$T \leq \frac{1}{\mu_i} \log\left(1 + \frac{\Lambda}{\mathcal{S}}\right) \tag{9.17}$$

*then $\mathbf{0}$ is globally asymptotically stable in (9.16). Condition (9.17) is equivalent to $\Lambda \geq \mathcal{S}\left(e^{\mu_i T} - 1\right)$.*

*Proof.* We know (see Appendix 9.A and Remark 9.3) that $M_i^{\text{crit}} \leq \frac{\mathcal{N}}{4\lambda\gamma_i}\left(1 - \frac{1}{\mathcal{N}}\right)^2$. Hence the following is a sufficient condition for global asymptotic stability of $\mathbf{0}$:

$$\underline{M}_i^{\text{imp}} \geq \frac{\mathcal{N}}{4\lambda\gamma_i}\left(1 - \frac{1}{\mathcal{N}}\right)^2 = \frac{(1-r)\,\nu_E \mathcal{N}}{4\mu_M \gamma_i}\left(1 - \frac{1}{\mathcal{N}}\right)^2 K.$$

That is

$$\frac{\Lambda e^{-\mu_i \tau}}{1-e^{-\mu_i \tau}} \geq \frac{(1-r)\,\nu_E \mathcal{N}}{4\mu_M \gamma_i}\left(1-\frac{1}{\mathcal{N}}\right)^2 K,$$

and the result is proved. □

**Remark 9.7.** *As a continuation of Remark 9.3, we note that Proposition 9.6 gives a very simple estimate for the target ratio of sterilizing males per release over initial wild male population as a function of the period between impulsive releases in the form*

$$\rho(T) := \frac{\Lambda}{M^*(M_i=0)} \simeq \left(e^{\mu_i T}-1\right)\frac{\mathcal{N}-1}{4\gamma_i}.$$

We can specify Lemma 9.5 for impulses and combine it with Proposition 9.3 to get a sufficient condition for extinction in the impulsive cases:

**Proposition 9.7.** *The impulsive control of amplitude $\Lambda > 0$ and period $T > 0$ satisfying $\Lambda \geq \mathcal{S}\left(e^{\mu_i T}-1\right)$ leads to extinction in $n$ impulses if*

$$n \geq \frac{\tau(\underline{M}_i^{imp})}{T}, \text{ where } \underline{M}_i^{imp} = \frac{\Lambda e^{-\mu_i T}}{1-e^{-\mu_i T}}. \tag{9.18}$$

## 9.3 Numerical study

### 9.3.1 Numerical method and parametrization

In order to preserve positivity of solutions and comparison principle, we use a nonstandard finite-differences (NSFD) scheme to integrate the differential systems (see for instance [10] for an overview).

For system (9.10), it reads

$$\begin{cases} \dfrac{E^{n+1}-E^n}{\Phi(\Delta t)} = bF_S^n(1-\dfrac{E^{n+1}}{K}) - (\nu_E+\mu_E)E^n, \\ \dfrac{M^{n+1}-M^n}{\Phi(\Delta t)} = (1-r)\nu_E E^n - \mu_M M^n, \\ \dfrac{M_i^{n+1}-M_i^n}{\Phi(\Delta t)} = -\mu_i M_i^n + u^n, \\ \dfrac{F^{n+1}-F^n}{\Phi(\Delta t)} = r\nu_E \dfrac{M^{n+1}}{M^{n+1}+M_i^{n+1}}\left(1-e^{-\beta\left(M^{n+1}+M_i^{n+1}\right)}\right)E^n - \mu_F F^n, \end{cases} \tag{9.19}$$

where $\Delta t$ is the time discretization parameter, $\Phi(\Delta t) = \frac{1-e^{-Q\Delta t}}{Q}$, $Q = \max\{\mu_M, \mu_F, \nu_E+\mu_E, \mu_i\}$ and $X^n$ (respectively $u^n$) is the approximation of $X(n\Delta t)$ (respectively $u(n\Delta t)$) for $n \in \mathbb{N}$.

| Parameter | $\beta$ | $b$ | $r$ | $\mu_E$ | $\nu_E$ | $\mu_F$ | $\mu_M$ | $\gamma_i$ | $\mu_i$ | $\Delta t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Value | $10^{-4}-1$ | 10 | 0.49 | 0.03 | $0.001-0.25$ | 0.04 | 0.1 | 1 | 0.12 | 0.1 |

Table 9.3: Numerical values fixed for the simulations.

We fix the value of some parameters using the values from Tables 9.1 and 9.2 (see Table 9.3). Then, in order to get results relevant for an island of 74 ha with an estimated male population of about 69 ha$^{-1}$, we let $\nu_E$ and $\beta$ vary, and fix $K$ such that

$$M_+^* = 69 \cdot 74 = 5106,$$

that is

$$K = \frac{5106 \cdot \mu_M}{(1-r)\nu_E\left(1-\frac{1}{\mathcal{N}\left(1-e^{-\beta \cdot 5106}\right)}\right)}.$$

Recall that for the choice from Table 9.3, page 172, we have

$$\mathcal{N} = 117.5\frac{\nu_E}{\nu_E+0.03}.$$

**Remark 9.8.** *Thus according to the values taken by* $\nu_E$ *in Table 9.3, page 172, we have the following bounds for* $\mathcal{N}$*:*

$$29 \leq \mathcal{N} \leq 105.$$

*The other aggregated value of interest,* $\psi = \frac{\mu_M}{(1-r)\nu_E \beta K} = \frac{\mathcal{N}-(1-e^{-\beta M_+^*})}{\mathcal{N} M_+^* \beta}$*, ranges from* $1.4 \cdot 10^{-4}$ *to* 2*, approximately.*

All computations were performed using Python programming language (version 3.6.2). The most costly operation was the separatrix approximation, which needed to be done once for each set of parameter values. We first compute points close to the separatrix (see details in Section 9.3.3), starting from a regular triangular mesh with 40 points on each side, then we reduce the points if any comparable pairs appeared. From these (at most 861) scattered points we build recursively a comparison tree by selecting the point $P$ which minimizes the distance to all other points, and distributing the remaining points into six subtrees, corresponding to each affine orthant whose vertex is $P$. Each tree was saved using `pickle` module, and loaded when necessary. This was done to reduce the number of operations for checking if a point is below the separatrix, as this needs to be done several times along each computed trajectory. Indeed, using the fact that two points on the separatrix cannot be related by the partial order, one only needs to investigate 3 of the 6 remaining orthants to determine if the candidate point is below any of the scattered points or not. For any given input of released sterilizing males, the computation of a trajectory ended either when the maximal number of iterations was reached (here, we fixed that value at $3 \cdot 10^5$) or when it was found below the separatrix, using the comparison tree. Trial CPU times (on a laptop computer with Intel® Core™ i5-2410M CPU @ 2.30GHz x 4 processor) for all these operations are given in Table 9.4.

| Operation | Approximation | Reduction | Tree building | Save | Load | Full trajectory | Stopped trajectory |
|---|---|---|---|---|---|---|---|
| CPU time (s) | 267 | 12 | 6.8 | $1.8 \cdot 10^{-3}$ | $1 \cdot 10^{-3}$ | 17 | 0.25 |

Table 9.4: CPU times for the numerical simulations

### 9.3.2 Equilibria and effort ratio

We first compute the position of equilibria for a range of values of $\beta$ and $\widetilde{\nu}_E$. This enables us to compute the effort ratio $\rho^*$, defined in Remark 9.3 as the ratio between the wild steady state male population $M^*(M_i = 0)$ and the critical constant value of sterilizing males $M_i^{\text{crit}}$ necessary in order to make $\mathbf{0}$ globally asymptotically stable. Values are shown in Table 9.5.

| $\nu_E$ | 0.005 | 0.010 | 0.020 | 0.030 | 0.050 | 0.100 | 0.150 | 0.200 | 0.250 |
|---|---|---|---|---|---|---|---|---|---|
| $\rho^*$ | 16 | 30 | 48 | 60 | 76 | 93 | 101 | 106 | 108 |

Table 9.5: Effort ratio $\rho^* = M_i^{\text{crit}}/M^*(M_i = 0)$ for various values of $\nu_E$. For this range of parameters, $\rho^*$ is practically independent on $\beta \in [10^{-4}, 1]$.

We note that $\rho^*$ depends practically only on $\nu_E$, because the Allee (with parameter $\beta$) does not apply at high population levels. In fact the ratio (and thus the control effort) increases with increasing values of $\nu_E$, that favor the maintenance of the wild population (the larger the value of $\nu_E$, the larger the value of $\mathcal{N}$ and the shorter the period in the eggs compartment).

### 9.3.3 Computation of the basin of attraction of 0 for (9.2)

We start from a regular triangular mesh of the triangle $\{(E, M, F) \in \mathbb{R}^3_+, \quad E + M + F = 1\}$, with 40 points on each side. Given $\epsilon > 0$, for each vertex $V$ of this mesh we compute $\lambda \in (0, +\infty)$ such that $\lambda V \in \Sigma_-$ and $(1+\epsilon)\lambda V \in \Sigma_+$. The points $\lambda V$ (which are numerically at distance at most $\epsilon$ of the separatrix $\Sigma$) are then plotted.

Figure 9.1 is typically the kind of figure that we can draw for each set of parameters. Depending on the parameters values, the basin of attraction of $\mathbf{0}$ can be tiny, or not. Its shape emphasizes the important role of eggs and, even, males abundance in the maintenance of the wild population. In fact, even if almost all females have disappeared, the control must go on in order to further reduce the stock of eggs before eventually reaching the separatrix.

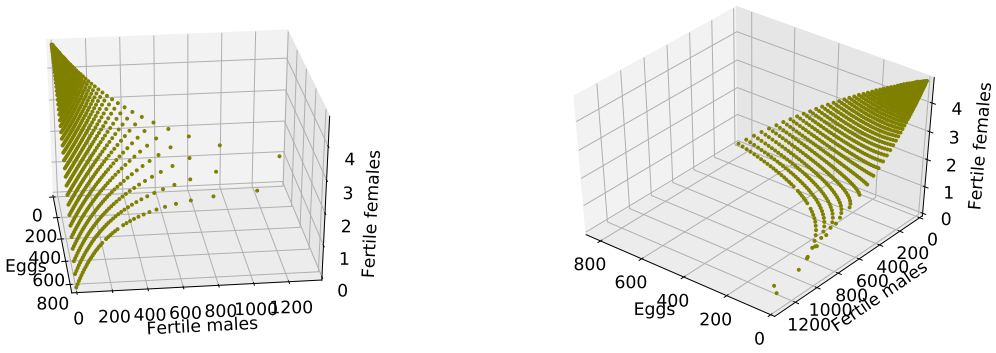



Figure 9.1: Two viewpoints of scattered points lying around the separatrix ($\epsilon = 10^{-2}$) for $\nu_E = 0.1$ and $\beta = 10^{-4}$. In this case, 5 females or 900 eggs are enough to prevent population elimination.

### 9.3.4 Constant releases and entrance time into basin

For the same set of parameters as before, we compute the entrance time into the basin of $\mathbf{0}$.

First, we use Proposition 9.3 to get in Table 9.6 an underestimation of the entrance time, whatever the releasing effort could be, these entrance times represent the minimal time under which the SIT control cannot be successful (in fact, this under-estimation corresponds to the situation where $M_i = +\infty$, that is an infinite releasing effort).

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.005 | 63 | 151 | 204 | 253 | 303 |
| 0.010 | 93 | 180 | 232 | 281 | 331 |
| 0.020 | 118 | 203 | 256 | 304 | 354 |
| 0.030 | 130 | 215 | 267 | 315 | 365 |
| 0.050 | 141 | 226 | 278 | 327 | 377 |
| 0.100 | 152 | 236 | 289 | 337 | 387 |
| 0.150 | 156 | 240 | 293 | 341 | 391 |
| 0.200 | 158 | 242 | 295 | 343 | 393 |
| 0.250 | 160 | 244 | 296 | 344 | 395 |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 258 | 351 | 448 | 545 | 642 |
| 286 | 374 | 464 | 553 | 643 |
| 301 | 381 | 462 | 544 | 625 |
| 307 | 383 | 461 | 538 | 615 |
| 332 | 404 | 477 | 550 | 623 |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 323 | 445 | 571 | 697 | 824 |
| 361 | 475 | 592 | 708 | 825 |
| 381 | 485 | 590 | 695 | 800 |
| 391 | 488 | 587 | 685 | 783 |
| 440 | 530 | 621 | 713 | 804 |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |

Table 9.6: Left: under-estimation of the entrance time into the basin of $\mathbf{0}$ from the analytic formula (9.8). Middle and right: over-estimation of the entrance time into the basin of $\mathbf{0}$ from formula (9.9) with $\epsilon = \frac{M_+^*}{M_+^* + \phi M_i^{\text{crit}}}$, when applicable, for $\phi = 8$ (middle) and $\phi = 4$ (right).

Then we compute numerically the entrance time for a range of releasing efforts. In details, computations were performed for $M_i = \phi M_i^{\text{crit}}$ with $\phi \in \{1.2, 1.4, 1.6, 1.8, 2, 4, 8\}$. Results are shown in Table 9.7 for $\phi = 1.2$, $\phi = 2$ and $\phi = 8$.

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.005 | 168 | 286 | 363 | 435 | 504 |
| 0.010 | 200 | 305 | 376 | 441 | 505 |
| 0.020 | 219 | 313 | 377 | 437 | 495 |
| 0.030 | 225 | 314 | 375 | 434 | 492 |
| 0.050 | 228 | 314 | 373 | 431 | 488 |
| 0.100 | 231 | 314 | 372 | 430 | 488 |
| 0.150 | 232 | 315 | 373 | 431 | 489 |
| 0.200 | 233 | 316 | 375 | 433 | 491 |
| 0.250 | 234 | 318 | 376 | 434 | 493 |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 148 | 262 | 338 | 409 | 478 |
| 180 | 283 | 352 | 417 | 480 |
| 199 | 292 | 355 | 415 | 473 |
| 207 | 295 | 355 | 413 | 471 |
| 212 | 297 | 355 | 413 | 470 |
| 215 | 298 | 356 | 414 | 472 |
| 217 | 300 | 358 | 416 | 474 |
| 219 | 302 | 360 | 418 | 476 |
| 220 | 303 | 362 | 420 | 478 |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 128 | 237 | 311 | 380 | 449 |
| 160 | 258 | 326 | 391 | 454 |
| 180 | 270 | 333 | 392 | 450 |
| 188 | 274 | 334 | 392 | 450 |
| 194 | 278 | 336 | 394 | 452 |
| 200 | 282 | 340 | 398 | 456 |
| 202 | 285 | 343 | 401 | 459 |
| 205 | 287 | 345 | 403 | 462 |
| 206 | 289 | 347 | 406 | 464 |

Table 9.7: Entrance time into the basin of $\mathbf{0}$ (in days) for various values of $(\nu_E, \beta)$, with $M_i = 1.2M_i^{\text{crit}}$ (left), $M_i = 2M_i^{\text{crit}}$ (middle) and $M_i = 8M_i^{\text{crit}}$ (right).

We notice that the entrance times corresponding to the biggest effort ratio are of the same order of magnitude as the analytic under-estimation from formula (9.8).

Another interesting output of Table 9.7 is that the release effort ratio is not so important in terms of duration of the control: depending on the values taken by $\nu_E$ and $\beta$, the lowest ratio

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.005 | 399 | 680 | 863 | 1034 | 1199 |
| 0.010 | 854 | 1302 | 1603 | 1880 | 2154 |
| 0.020 | 1513 | 2166 | 2603 | 3022 | 3423 |
| 0.030 | 1950 | 2726 | 3253 | 3761 | 4264 |
| 0.050 | 2482 | 3421 | 4059 | 4686 | 5315 |
| 0.100 | 3100 | 4218 | 5000 | 5777 | 6553 |
| 0.150 | 3383 | 4581 | 5434 | 6278 | 7122 |
| 0.200 | 3545 | 4806 | 5694 | 6578 | 7461 |
| 0.250 | 3649 | 4956 | 5869 | 6779 | 7689 |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 587 | 1036 | 1338 | 1619 | 1893 |
| 1283 | 2009 | 2499 | 2962 | 3416 |
| 2296 | 3367 | 4092 | 4782 | 5452 |
| 2989 | 4260 | 5132 | 5976 | 6815 |
| 3837 | 5381 | 6434 | 7483 | 8529 |
| 4817 | 6675 | 7975 | 9268 | 10563 |
| 5274 | 7268 | 8688 | 10095 | 11502 |
| 5549 | 7638 | 9117 | 10588 | 12060 |
| 5717 | 7884 | 9405 | 10922 | 12438 |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 2024 | 3749 | 4920 | 6027 | 7115 |
| 4548 | 7343 | 9257 | 11112 | 12912 |
| 8317 | 12451 | 15331 | 18040 | 20736 |
| 10862 | 15871 | 19319 | 22691 | 26040 |
| 14058 | 20188 | 24395 | 28588 | 32774 |
| 17891 | 25266 | 30457 | 35640 | 40813 |
| 19651 | 27618 | 33285 | 38913 | 44541 |
| 20757 | 29073 | 34979 | 40865 | 46750 |
| 21443 | 30036 | 36123 | 42188 | 48254 |

Table 9.8: Total effort ratio to get into the basin of $\mathbf{0}$ for various values of $(\nu_E, \beta)$, with $M_i = 1.2M_i^{\text{crit}}$ (left), $M_i = 2M_i^{\text{crit}}$ (middle) and $M_i = 8M_i^{\text{crit}}$ (right). The total effort ratio in this case is defined as $M_i/M_+^*$ multiplied by $\mu_i$ times the entrance time, and corresponds to the number of males that should be released at a constant level, divided by the initial male population.

needs between 4 to 7 more weeks to reach the basin, than the largest ratio. Contrary to what could have been expected, there is no linear relationship. This can be explained by the fact that a female mates only once. Thus if males are in abundance, all females have mated, and then many released males become useless with regards to sterilization. Of course, this has to be mitigated taking into account that our model implicitly assumes a homogeneous distribution, while in real, environmental parameters (like vegetation, climate, etc.) have to be taken into account [72]. Last but not least, Table 9.8, page 175, clearly emphasizes that a large effort ratio, *i.e.* $\phi = 8$, means the use (and then the production) of a large number of sterile males with a really small time-saving compared to the case $\phi = 2$. For instance with $\nu_E = 0.05$ and $\beta = 10^{-2}$, the total effort ratio for $\phi = 8$ is approximately 6 times larger than for $\phi = 2$ (24395 against 4059), with a time-saving of 37 days, that is approximately one tenth of the total protocol duration (336 days against 373).

In other words, releasing a large number of sterile males is not necessarily a good strategy, from the economical point of view, but also from the control point of view.

In the next subsection, We consider a more realistic scenario, where sterile males are released periodically and instantaneously (system (9.16)).

### 9.3.5 Periodic releases

In the case of periodic releases by pulses $u = [T, \Lambda\delta_0, \infty]$, for a given couple $(\nu_E, \beta)$ we compute the first time $t > 0$ such that $(E, M, F)(t)$ is below one point of the previously computed separatrix.

We performed the computations with $T \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, choosing

$$\Lambda = K\frac{\phi(1-r)\nu_E\mathcal{N}}{4\mu_M}\big(1-\frac{1}{\mathcal{N}}\big)^2\big(e^{\mu_i T}-1\big)$$

for $\phi \in \{1.2, 1.4, 1.6, 1.8, 2, 4, 8\}$.

For all combinations of $(\nu_E, \beta)$, we indicate in Table 9.9 the maximal and minimal (with respect to $(T, \phi)$) total effort ratio $\rho_{\text{tot}}$ defined as the number of released mosquitoes at the time when the basin of $\mathbf{0}$ is reached, divided by the initial male population that is:

$$\rho_{\text{tot}} := n_{\text{tot}}\Lambda/M_+^*, \quad n_{\text{tot}} = \min\{\lfloor t/T \rfloor, (E, M, F)(t) \in \Sigma_-\}.$$

These extremal values are obtained for a period $T$ and with an entrance time $t_*$ that are shown in parentheses. We also indicate in Table 9.10 the maximal and minimal entrance times, obtained for a period $T$ and an effort ratio $\rho_{\text{tot}}$ that are shown in parentheses. Note that consistently, the minimal entrance time is always obtained for $\phi = 8$ and corresponds to the maximal effort ratio. Maximal entrance time is obtained for $T = 1$ (minimal tested period) and the minimal entrance time is obtained for $T = 10$ (maximal tested period). However, the minimal effort ratio is sometimes obtained with $T = 2$.

Comparing Tables 9.8 and 9.9 shows that in general, a periodic control achieves the target of bringing the population into $\Sigma_-$ at a smaller cost than the constant control (in terms of total number of released mosquitoes, counted with respect to the wild population).

### 9.3.6 Case study: Onetahi motu

We now parametrize explicitly our model to the case of Onetahi motu in Tetiaroa atoll (French Polynesia), where weekly ($T = 7$ days) releases have been performed over a year. Male population

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.005 | 282 (2, 287) | 384 (2, 491) | 448 (1, 608) | 502 (1, 682) | 554 (1, 752) |
| 0.010 | 547 (1, 344) | 698 (2, 497) | 796 (1, 602) | 884 (1, 669) | 969 (2, 805) |
| 0.020 | 900 (1, 357) | 1112 (1, 519) | 1253 (1, 585) | 1386 (1, 647) | 1504 (2, 771) |
| 0.030 | 1125 (3, 363) | 1371 (1, 510) | 1538 (1, 572) | 1696 (2, 693) | 1839 (2, 752) |
| 0.050 | 1383 (2, 379) | 1669 (1, 496) | 1875 (1, 556) | 2066 (2, 672) | 2238 (2, 730) |
| 0.100 | 1655 (2, 370) | 1997 (1, 480) | 2238 (1, 539) | 2458 (2, 650) | 2678 (2, 708) |
| 0.150 | 1772 (1, 388) | 2134 (1, 473) | 2394 (2, 583) | 2632 (2, 641) | 2871 (2, 699) |
| 0.200 | 1834 (1, 384) | 2213 (1, 470) | 2482 (2, 578) | 2731 (2, 636) | 2979 (1, 738) |
| 0.250 | 1873 (1, 382) | 2263 (1, 468) | 2531 (2, 575) | 2787 (2, 633) | 3043 (2, 692) |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 1095 (10, 135) | 1838 (10, 248) | 2450 (10, 323) | 2986 (10, 393) | 3522 (10, 462) |
| 2317 (10, 168) | 3575 (10, 268) | 4536 (10, 337) | 5499 (10, 402) | 6323 (10, 466) |
| 4139 (10, 188) | 6015 (10, 280) | 7573 (10, 343) | 8909 (10, 402) | 10246 (10, 460) |
| 5448 (10, 196) | 7829 (10, 283) | 9506 (10, 343) | 11183 (10, 402) | 12581 (10, 460) |
| 7155 (10, 201) | 9818 (10, 286) | 11921 (10, 344) | 14025 (10, 402) | 15778 (10, 460) |
| 8794 (10, 206) | 12114 (10, 289) | 14709 (10, 347) | 17305 (10, 405) | 19900 (10, 463) |
| 9522 (10, 209) | 13603 (10, 291) | 15948 (10, 350) | 18762 (10, 408) | 21576 (10, 466) |
| 10431 (10, 211) | 14201 (10, 293) | 17138 (10, 352) | 19586 (10, 410) | 22524 (10, 468) |
| 10709 (10, 212) | 14584 (10, 295) | 17601 (10, 353) | 20618 (10, 412) | 23133 (10, 470) |

Table 9.9: Minimal (left) and maximal (right) total effort ratio to get into the basin of **0** (in days) for various values of $(\nu_E, \beta)$, the minimum and maximum being taken with respect to $(T, \phi)$, with a period and an entrance time shown in parentheses. The total effort ratio is defined as the total number of released male mosquitoes divided by the initial (wild) male mosquito population.

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.005 | 135 (10, 1095) | 248 (10, 1838) | 323 (10, 2450) | 393 (10, 2986) | 462 (10, 3522) |
| 0.010 | 168 (10, 2317) | 268 (10, 3575) | 337 (10, 4536) | 402 (10, 5499) | 466 (10, 6323) |
| 0.020 | 188 (10, 4139) | 280 (10, 6015) | 343 (10, 7573) | 402 (10, 8909) | 460 (10, 10246) |
| 0.030 | 196 (10, 5448) | 283 (10, 7829) | 343 (10, 9506) | 402 (10, 11183) | 460 (10, 12581) |
| 0.050 | 201 (10, 7155) | 286 (10, 9818) | 344 (10, 11921) | 402 (10, 14025) | 460 (10, 15778) |
| 0.100 | 206 (10, 8794) | 289 (10, 12114) | 347 (10, 14709) | 405 (10, 17305) | 463 (10, 19900) |
| 0.150 | 209 (10, 9522) | 291 (10, 13603) | 350 (10, 15948) | 408 (10, 18762) | 466 (10, 21576) |
| 0.200 | 211 (10, 10431) | 293 (10, 14201) | 352 (10, 17138) | 410 (10, 19586) | 468 (10, 22524) |
| 0.250 | 212 (10, 10709) | 295 (10, 14584) | 353 (10, 17601) | 412 (10, 20618) | 470 (10, 23133) |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 456 (1, 317) | 667 (1, 420) | 752 (1, 474) | 826 (1, 521) | 896 (1, 565) |
| 528 (1, 629) | 661 (1, 749) | 735 (1, 833) | 803 (1, 909) | 868 (1, 982) |
| 534 (1, 1012) | 642 (1, 1179) | 708 (1, 1300) | 771 (1, 1414) | 830 (1, 1522) |
| 527 (1, 1246) | 627 (1, 1445) | 690 (1, 1588) | 749 (1, 1724) | 807 (1, 1860) |
| 514 (1, 1513) | 605 (1, 1749) | 666 (1, 1925) | 724 (1, 2090) | 782 (1, 2257) |
| 494 (1, 1787) | 581 (1, 2072) | 640 (1, 2279) | 698 (1, 2485) | 755 (1, 2692) |
| 483 (1, 1896) | 569 (1, 2200) | 628 (1, 2428) | 686 (1, 2652) | 744 (1, 2877) |
| 477 (1, 1953) | 563 (1, 2272) | 622 (1, 2510) | 680 (1, 2745) | 738 (1, 2979) |
| 473 (1, 1988) | 559 (1, 2317) | 618 (1, 2562) | 676 (1, 2802) | 734 (1, 3043) |

Table 9.10: Minimal (left) and maximal (right) entrance time into the basin of **0** (in days) for various values of $(\nu_E, \beta)$, the minimum and maximum being taken with respect to $(T, \phi)$, with a period and a total effort ratio shown in parentheses.

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.001 | 39 | 200 | 295 | 376 | 453 |
| 0.002 | 142 | 310 | 402 | 480 | 555 |
| 0.005 | 877 | 1094 | 1178 | 1252 | 1323 |
| 0.008 | N/A | N/A | N/A | N/A | N/A |
| 0.010 | N/A | N/A | N/A | N/A | N/A |
| 0.015 | N/A | N/A | N/A | N/A | N/A |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 34 | 181 | 272 | 352 | 430 |
| 111 | 262 | 350 | 428 | 503 |
| 350 | 471 | 554 | 627 | 697 |
| 1167 | 1091 | 1168 | 1238 | 1305 |
| N/A | N/A | N/A | N/A | N/A |
| N/A | N/A | N/A | N/A | N/A |

| $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|
| 30 | 171 | 261 | 341 | 418 |
| 97 | 241 | 327 | 404 | 480 |
| 260 | 381 | 462 | 535 | 605 |
| 443 | 541 | 618 | 687 | 754 |
| 676 | 728 | 802 | 870 | 935 |
| N/A | N/A | N/A | N/A | N/A |

Table 9.11: Entrance time into the basin of **0** (in days) for various values of $(\nu_E, \beta)$ with constant weekly ($T = 7$ days) releases at $p = 4$ (left), $p = 6$ (center) or $p = 8$ (right).

was estimated at $69 \cdot 74 \simeq 5000$ individuals, and the initial effort ratio $p := \Lambda / M_+^*$ was estimated at 8.

| $\nu_E \backslash \beta$ | $10^{-4}$ | $10^{-3}$ | $10^{-2}$ | $10^{-1}$ | $10^{0}$ |
|---|---|---|---|---|---|
| 0.001 | 0.943252 | 0.147678 | 0.020134 | 0.002495 | 0.000283 |
| 0.002 | 0.567382 | 0.071552 | 0.009875 | 0.001247 | 0.000141 |
| 0.005 | 0.205116 | 0.031070 | 0.004439 | 0.000568 | 0.000069 |
| 0.008 | 0.133889 | 0.021388 | 0.003170 | 0.000425 | 0.000052 |
| 0.010 | 0.111803 | 0.018284 | 0.002779 | 0.000380 | 0.000047 |
| 0.015 | N/A | N/A | N/A | N/A | N/A |

Table 9.12: Final total female ratio $\frac{(F+F_{st})(t)}{F^*+F^*_{st}}$ at time $t$ when the trajectory enter into the basin of **0** for various values of $(\nu_E, \beta)$ with constant weekly ($T = 7$ days) releases at $p = 8$.

For $p \in \{4, 6, 8\}$, entrance times (in days) are shown in Table 9.11 and final total female ratio in Table 9.12. This last quantity is important for practical purposes to help answering the question: when is it time to stop the releases? The trap counts during the experiment are to be compared with the initial trap counts (before the releases), and roughly, the process can be stopped once the ratio between the counts goes below the values in Table 9.12. Interestingly, $\beta$ determines the order of magnitude of this final ratio.

Table 9.11 provides us interesting information on the entrance time versus the transition rate $\nu_E$ and the mating parameter $\beta$. If the effort ratio $p$ is not large enough, the SIT treatment can fail, and even if it is large enough (say $p = 8$) the time to reach the basin of **0** can be very large.

In the 3-dimensional state space $(E, M, F)$ we draw the full trajectory for the same sample value ($\nu_E = 0.008$, $\beta = 10^{-3}$, $p = 8$) along with a zoom in the last 30 days of treatment showing also the separatrix between the basins of $\mathbf{E}_+$ and **0** as dots in Figure 9.2. According to Table 9.11,

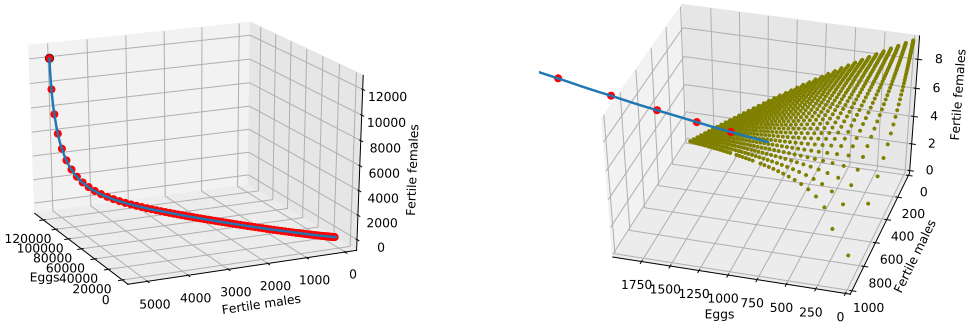
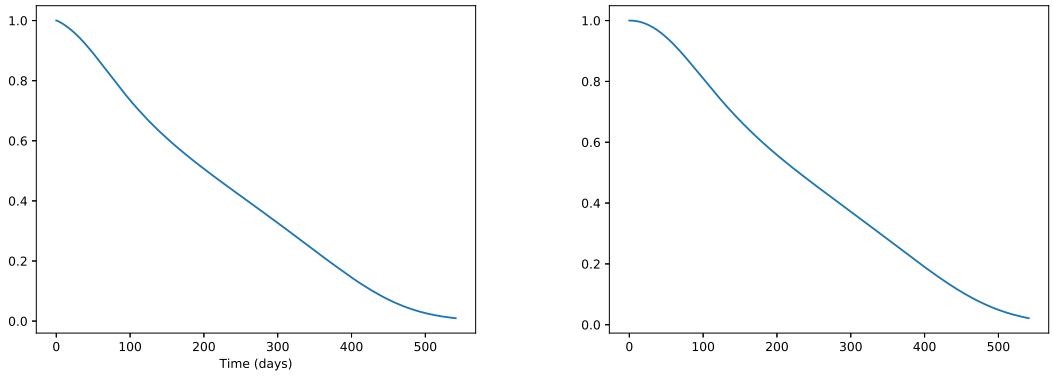
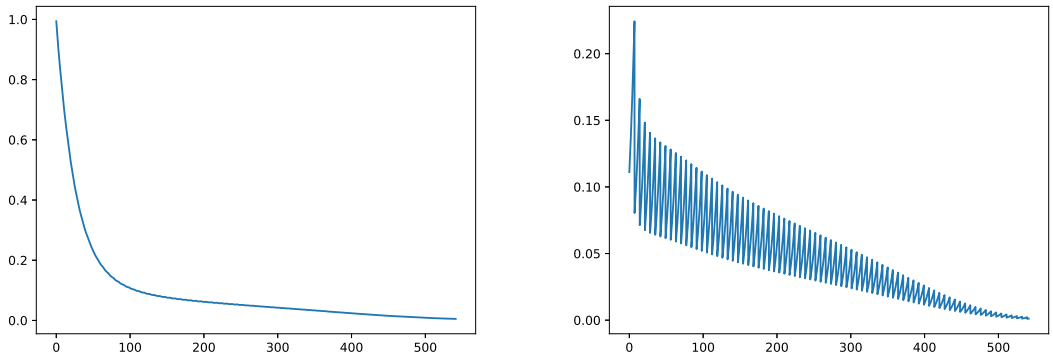



Figure 9.2: Trajectory $t \mapsto (E(t), M(t), F(t))$ for $\nu_E = 0.008$ and $\beta = 10^{-3}$ (left) and a zoom in the last 30 days of treatment displaying also the separatrix as dots (right).




(a) Relative egg count $\frac{E(t)}{E^*_+}$.

(b) Relative total female count $\frac{F(t)+F_{st}(t)}{F^*_+ + F^*_{st,+}}$.



(c) Fertile female ratio $\frac{F(t)}{F(t)+F_{st}(t)}$.

(d) Fertile male ratio $\frac{M(t)}{M(t)+M_i(t)}$.

Figure 9.3: Time dynamics of different ratios for $\nu_E = 0.008$ and $\beta = 10^{-3}$.

page 176, the entrance time is 541, which justifies that the control should last for more than one year. Our system being monotone, the trajectory is monotone decreasing (see Figure 9.2 (left), page 177). However, the rate of the decrease is relatively large at the beginning of the treatment, and then becomes small and, almost, constant. We also show time dynamics of four relevant normalized quantities, for the same sample value ($\nu_E = 0.008$, $\beta = 10^{-3}$, $p = 8$) in Figure 9.3.

## 9.4 Conclusion

In this paper we have derived a minimalistic model to control mosquito population by Sterile Insect Technique, using either irradiation or the cytoplasmic incompatibility of *Wolbachia* to release sterilizing males. We particularly focus on the chance of collapsing the wild population, provided that the selected area allows elimination. Thus contrary to previous SIT and IIT models, the trivial equilibrium, $\mathbf{0}$ is always Locally Asymptotically Stable, at least. We consider different type of releases (constant, continuous, or periodic and instantaneous) and show necessary conditions to reach eradication, in each case. We also derived the minimal time under which eradication cannot occur, (*i.e.* entrance into the basin of attraction of $\mathbf{0}$ is impossible), whatever the control effort. Obviously, the knowledge on the mosquito parameters are very important, particularly the duration of the egg compartment, $\dfrac{1}{\mu_E + \nu_E}$ and the mating parameter, $\beta$. Surprisingly, mosquito entomologists have not yet really focused their experiments on $\beta$ or the probability of meeting/mating between one male and one female according to the size of the domains. Our model illustrates the importance of this parameter (and others) in the duration of the SIT control. In general, SIT entomologists recommend to release a minimum of ten times more sterile males than (estimated) wild males: this can be necessary if the competitiveness of the sterile male is weak compared to the wild ones (this can be the case with irradiation SIT approach). Our approach may help standardizing and quantifying this estimated ratio.

Finally, we focus on a real case, the Onetahi motu, where a *Wolbachia* experiment has been conducted by Dr. Hervé Bossin and his collaborators, driving the local mosquito, *Aedes polynesiensis*, to nearly elimination. Our preliminary results show some good agreement with field observation (mainly trapping).

Our results also show the importance of eggs in the survival of the wild population. If the egg stock is sufficiently large, and depending on weather parameters, the wild population can re-emerge after the control has stopped. That is why, according to our model and numerical results, it is recommended to pursue the release of sterilizing males even after wild mosquito females are no longer collected in monitoring traps.

Last but not least, we hope that our theoretical results will be helpful to improve future SIT experiments and particularly to take into account the long term dynamics of eggs.

# Appendices

## 9.A Study of the steady states

This section is devoted to the proof of Lemma 9.3.

### 9.A.1 Study of $f$

We first study function $f$ defined in (9.4). For any $y \geq 0$, if $x \geq \frac{1}{\psi}$ then $f(x,y) < -\frac{1}{\mathcal{N}}(x+y)$ so in particular $f(x,y) < 0$. Therefore all steady states must satisfy $\beta M^* < \frac{1}{\psi}$. Likewise,

$$y \geq 0,\ 0 \leq x < \frac{1}{\psi} \implies (1-\psi x)(1-e^{-(x+y)}) < 1.$$

Hence for all $x < \frac{1}{\psi}$ we find $f(x,y) < (1-\frac{1}{\mathcal{N}})x - \frac{1}{\mathcal{N}}y$. As a consequence, if $\mathcal{N} \leq 1$ then $f(x,y) < 0$ for all $(x,y) \in \mathbb{R}_+^2 \backslash \{0\}$, and system (9.2) has no positive steady state. From now on we assume that $\mathcal{N} > 1$.

We also compute directly $f(0,y) = -\frac{1}{\mathcal{N}}y < 0$ and $\lim_{x\to+\infty} f(x,y) = -\infty$.

**Remark 9.9.** *For all $x \in (0, 1/\psi)$, we notice that*

$$f(x,y) < Q_y(x) = -\psi x^2 + (1-\frac{1}{\mathcal{N}})x - \frac{y}{\mathcal{N}}.$$

*The discriminant of the second-order polynomial $Q_y$ is*

$$\Delta_y = (1-\frac{1}{\mathcal{N}})^2 - \frac{4y\psi}{\mathcal{N}}.$$

*Let $\widetilde{y} := \frac{\mathcal{N}}{4\psi}(1-\frac{1}{\mathcal{N}})^2$. If $y \geq \widetilde{y}$ then $\Delta_y \leq 0$, hence $f < 0$. At this stage we know that if $\beta\gamma_i M_i \geq \widetilde{y}$ then there is no positive steady state.*

*The quantity $\widetilde{y}$ is used in Remark 9.3 to obtain a first-order approximation of the target release ration.*

We now compute the derivatives of $f$:

$$\begin{aligned}
\partial_x f &= (1-2\psi x)(1-e^{-(x+y)}) - \frac{1}{\mathcal{N}} + x(1-\psi x)e^{-(x+y)},\\
\partial^2_{xx} f &= -2\psi(1-e^{-(x+y)}) + e^{-(x+y)}\big(2-(4\psi+1)x+\psi x^2\big)\\
\partial^3_{xxx} f &= e^{-(x+y)}\Big(-6\psi-3+(6\psi+1)x-\psi x^2\Big) =: e^{-(x+y)}Q_3(x)\\
\partial_y f &= x(1-\psi x)e^{-(x+y)} - \frac{1}{\mathcal{N}},\\
\partial^2_{yy} f &= -x(1-\psi x)e^{-(x+y)} < 0 \text{ for } x \in (0,1/\psi).
\end{aligned}$$

Obviously, $\partial_x f(x,y) < 0$ if $x \geq \frac{1}{\psi}$ and $\partial_x f(0,y) = 1 - e^{-y} - \frac{1}{\mathcal{N}}$, which is positive if and only if $y > -\log(1-\frac{1}{\mathcal{N}}) = \log(1+\frac{1}{\mathcal{N}-1})$.

In order to know the variations of $\partial^3_{xxx} f$ we study the second-order polynomial

$$Q_3(x) = -6\psi - 3\beta + x\big(6\psi+1\big) - \psi x^2.$$

Its discriminant is

$$\Delta_3 = (6\psi+1)^2 - 4\psi(6\psi+3) = 1+12\psi^2,$$

which is positive. Therefore $\partial^3_{xxx} f$ is negative-positive-negative. More precisely, $Q_3$ is positive on

$$(w_-, w_+) := \big(\frac{6\psi+1-\sqrt{1+12\psi^2}}{2\psi}, \frac{6\psi+1+\sqrt{1+12\psi^2}}{2\psi}\big).$$

To go one step further, we need to know the signs of $\partial^2_{xx} f(w_+, y)$ and $\partial^2_{xx} f(0,y)$. We write

$$\partial^2_{xx} f(x,y) > 0 \iff e^{-(x+y)}\Big(2+2\psi-(4\psi+1)x+\psi x^2\Big) > 2\psi$$

Hence $\partial^2_{xx} f(0,y) > 0$ if and only if $y < \log(1+\frac{1}{\psi})$. Similarly, $\partial^2_{xx} f(w_+, y) < 0$ if and only if

$$y > \log\big(1+\frac{1}{\psi} - (2+\frac{1}{2\psi})w_+ + \frac{1}{2}w_+^2\big) - w_+.$$

This is always true:

**Lemma 9.6.** *For all $\psi > 0$,*

$$\log\big(1+\frac{1}{\psi} - (2+\frac{1}{2\psi})w_+ + \frac{1}{2}w_+^2\big) - w_+ < 0.$$

*Proof.* To prove it, we introduce $\gamma = \frac{1}{2\psi}$ so that we are left with

$$\log\big(7+3\gamma+\gamma^2+(4+\gamma)\sqrt{3+\gamma^2}\big) - (3+\gamma+\sqrt{3+\gamma^2}) < 0.$$

To check this we introduce

$$g(x) := \log(7+3x+x^2+(4+x)\sqrt{3+x^2}) - (3+x+\sqrt{3+x^2}),$$

and we want to prove that $g$ is negative. We compute that the sign of $g'(x)$ is equal to that of

$$-(4+x)(3+x^2) - 2x - \sqrt{3+x^2}(8+2x+x^2) < 0.$$

It remains to check that $g(0) = \log(7+4\sqrt{3}) - (3+\sqrt{3}) < 0$, which is true since

$$e^{3+\sqrt{3}} > e^4 > 2^4 > 7+8 > 7+4\sqrt{3},$$

where we used $e > 2$ and $1 < \sqrt{3} < 2$. □

Thus we obtain that $x \mapsto \partial^2_{xx} f(x,y)$ is either positive-negative (if $y < \log(1+\frac{1}{\psi})$) or always negative (otherwise).

The conclusion of all these computations is that in both cases ($f$ is either convex-concave or simply concave), for any $y$, $f(0,y) < 0$, $f(+\infty, y) = -\infty$ so that all in all there are either 0, 1 or 2 solutions to $f(x,y) = 0$, depending merely on the sign of the maximum of $x \mapsto f(x,y)$.

### 9.A.2 Study of functions $h_\pm$

We move on to the next step of the proof, studying the functions $h_\pm$ defined in (9.5). Recall that solving $f(x,y) = 0$ (for $x, y > 0$) is equivalent to picking $\theta = e^{-(x+y)} \in (0,1)$ and $y = h_\pm(\theta)$.

First, to check that $h_+$ and $h_-$ are well-defined we need to check that $1 + \xi\frac{\log(\theta)}{1-\theta} > 0$ for some $\theta \in (0,1)$. It is easily checked that this is the case on $(\theta_0(\xi), 1)$, and $\theta_0(\xi)$ is well-defined as soon as $\xi < 1$.

Hence if $\xi \geq 1$ then there is no nonzero steady state. Assume therefore that $\xi < 1$. Then there exists a unique $\theta_0(\xi) \in (0,1)$ such that $1-\theta-\xi\log(\theta)$ has the same sign as $\theta - \theta_0$ on $(0,1)$, that is, $1-\theta_0 = \frac{4\psi}{\mathcal{N}}\log(\theta_0)$.

We can check that $h_-$ is decreasing, $h_- < h_+$ on $(\theta_0, 1]$,

$$h_\pm(\theta_0) = -\frac{1}{2\psi} - \log(\theta_0),$$

and

$$h_-(1) < h_+(1) = \frac{1}{2\psi}\big(-1+\sqrt{1-\xi}\big) < 0.$$

Indeed (recall that $\mathcal{N}\xi = 4\psi$),

$$h'_-(\theta) = -\frac{1}{\theta} - \frac{1}{\mathcal{N}}\frac{\frac{1}{\theta(1-\theta)} + \frac{\log(\theta)}{(1-\theta)^2}}{\sqrt{1+\xi\frac{\log(\theta)}{1-\theta}}} < 0,$$

since

$$-\frac{\log(\theta)}{1-\theta} < \frac{1}{\theta}.$$

Let $y^{\text{crit}} := \max_{\theta\in[\theta_0(\xi),1]} h_+(\theta)$. If $y = y^{\text{crit}}$ then there is exactly one solution to $f(x,y) = 0$. For any $y \in [0, y^{\text{crit}})$, there are at least two solutions. By the previous computations we know that there are at most two solutions. So in this case there are exactly two solutions. To describe them one should consider $I_1 := [0, h_-(\theta_0(\xi))]$, if $h_-(\theta_0(\xi)) > 0$ ($I_1 = \emptyset$ otherwise), and $I_2 = (\max(\theta_0(\xi), 0), y^{\text{crit}})$. If $y \in I_1$ then there is a solution of the form $h_-(\theta_-)$ and one of the form $h_+(\theta_-)$. If $y \in I_2$ then both solutions are of the form $h_+(\theta)$, for two values of $\theta$ whose range contains the argument of $y^{\text{crit}}$. And for $y > y^{\text{crit}}$ there is no solution.

At this stage we proved that if $\xi \geq 1$ then there is no positive steady state; if $\xi < 1$ then if $y^{\text{crit}} > 0$ then there are two positive steady states for $\beta\gamma_i M_i \in [0, y^{\text{crit}})$, 1 for $\beta\gamma_i M_i = y^{\text{crit}}$ and 0 for $\beta\gamma_i M_i > y^{\text{crit}}$. If $y^{\text{crit}} = 0$ then there is a unique positive steady state and if $y^{\text{crit}} < 0$ then there is no positive steady state for any $M_i \geq 0$.

### 9.A.3 Stability

Finally, in order to compute the linearized stability of the steady states, we decompose $J = M_0 + N_0$, where $M_0$ is non-negative and $N_0$ is diagonal non-positive. Then $J$ (being Metzler, since $E < K$ at steady states) is stable if and only if $\rho(-N_0^{-1}M_0) < 1$. We compute

$$N_0 = \begin{pmatrix} -\frac{bF}{K} - (\nu_E + \mu_E) & 0 & 0 \\ 0 & -\mu_M & 0 \\ 0 & 0 & -\mu_F \end{pmatrix}$$

and

$$M_0 = \begin{pmatrix} 0 & 0 & b(1-\frac{E}{K}) \\ (1-r)\nu_E & 0 & 0 \\ \frac{r\nu_E M}{M+\gamma_i M_i}(1-e^{-\beta(M+\gamma_i M_i)}) & \frac{r\nu_E E}{M+\gamma_i M_i}\big(\beta M e^{-\beta(M+\gamma_i M_i)} + \frac{\gamma_i M_i}{M+\gamma_i M_i}(1-e^{-\beta(M+\gamma_i M_i)})\big) & 0 \end{pmatrix}$$

so that for some $X_1, X_2 \in \mathbb{R}$ (which we compute below at steady states) we have

$$-N_0^{-1}M_0 = \begin{pmatrix} 0 & 0 & \frac{b(1-\frac{E}{K})}{b\frac{F}{K}+\nu_E+\mu_E} \\ \frac{(1-r)\nu_E}{\mu_M} & 0 & 0 \\ \frac{X_1}{\mu_F} & \frac{X_2}{\mu_F} & 0 \end{pmatrix}.$$

At the steady state $(0,0,0)$, we have directly unconditional stability as

$$J = \begin{pmatrix} -(\nu_E+\mu_E) & 0 & b \\ (1-r)\nu_E & -\mu_M & 0 \\ 0 & 0 & -\mu_F \end{pmatrix},$$

whose eigenvalues are $-(\nu_E+\mu_E)$, $-\mu_M$ and $-\mu_F$.

At a non-zero steady state we recall that

$$\begin{aligned} bF &= \frac{(\nu_E+\mu_E)E}{1-\frac{E}{K}}, \\ E &= \lambda K M, \\ r\nu_E(1-e^{-\beta(M+\gamma_i M_i)})\frac{M}{M+\gamma_i M_i} &= \mu_F\frac{F}{E} = \frac{\mu_F(\nu_E+\mu_E)}{b}\frac{1}{1-\lambda M}, \\ e^{-\beta(M+\gamma_i M_i)} &= 1 - \frac{1}{\mathcal{N}(1-\lambda M)}\frac{M+\gamma_i M_i}{M}, \end{aligned}$$

so that

$$X_1 = \frac{r\nu_E}{\mathcal{N}(1-\lambda M)},$$
$$X_2 = \frac{r\nu_E \lambda K M}{M+\gamma_i M_i}\Big(\beta M\big(1-\frac{M+\gamma_i M_i}{\mathcal{N}M(1-\lambda M)}\big)+\frac{\gamma_i M_i}{M}\frac{1}{\mathcal{N}(1-\lambda M)}\Big).$$

The characteristic polynomial of $-N_0^{-1}M_0$ is

$$P(z) = -z^3 + \frac{b(1-\lambda M)^2}{\nu_E+\mu_E}\Big(\frac{(1-r)\nu_E X_2}{\mu_M\mu_F} + z\frac{X_1}{\mu_F}\Big),$$

which is equal to

$$P(z) = -z^3+\mathcal{N}(1-\lambda M)^2\Big(\frac{M}{M+\gamma_i M_i}\big(\beta M(1-\frac{M+\gamma_i M_i}{\mathcal{N}M(1-\lambda M)})+\frac{\gamma_i M_i}{M\mathcal{N}(1-\lambda M)}\big)+\frac{z}{\mathcal{N}(1-\lambda M)}\Big),$$

and we rewrite it as

$$P(z) = -z^3 + (1-\lambda M)\Big(\beta\mathcal{N}\frac{M^2(1-\lambda M)}{M+\gamma_i M_i} - \beta M + \frac{\gamma_i M_i}{M+\gamma_i M_i} + z\Big)$$

We find $P(0) > 0$ (since $X_2 > 0$) and

$$P'(z) = -3z^2 + (1-\lambda M),$$

so that $J$ is stable if and only if $P(1) < 0$. ($P$ is increasing and then decreasing on $(0, +\infty)$). This condition reads

$$(1-\lambda M)\Big(1+\frac{\gamma_i M_i}{M+\gamma_i M_i} + \beta M\big(-1+\mathcal{N}\frac{M}{M+\gamma_i M_i}(1-\lambda M)\big)\Big) < 1. \tag{9.20}$$

Let us treat first the case when $M_i = 0$. The stability condition rewrites

$$(1-\lambda M)\big(1+\beta M(-1+\mathcal{N}(1-\lambda M))\big) < 1,$$

that is, for a nonzero steady state,

$$-\lambda + \beta(-1+\mathcal{N}(1-\lambda M)) - \lambda\beta M(-1+\mathcal{N}(1-\lambda M)) < 0.$$

If $M_i^{\text{crit}} > 0$, we know that there are exactly two steady states between 0 and $1/\lambda$ for $M_i = 0$, which we denote by $0 < M_- < M_+ < 1/\lambda$. Let $\phi(x) = 1 - \frac{1}{\mathcal{N}} - \lambda x + e^{-\beta x}(\lambda x - 1)$. We have $\phi(M_\pm) = 0$ and $\mp\phi'(M_\pm)) > 0$.

In particular, $\phi'(M_+) > 0$ so

$$M_+ > \frac{1}{\lambda} + \frac{1}{\beta}(1-e^{\beta M_+}) = \frac{1}{\lambda} - \frac{1}{\beta}\frac{1}{(1-\lambda M_+)\mathcal{N}-1}.$$

Multiplying this inequality by $\lambda\beta\big((1-\lambda M_+)\mathcal{N}-1\big)$ yields exactly the stability of $M_+$, since $(1-\lambda M_+)\mathcal{N} > 1$. Indeed,

$$\mathcal{N}(1-\lambda M_\pm) = \frac{e^{\beta M_\pm}}{e^{\beta M_\pm}-1} > 1.$$

By a similar computation one can show that the smaller steady state $M_-$ is unstable.

We move now to the general case $M_i \geq 0$, assume $M_i < M_i^{\text{crit}}$ and write that $\partial_x f < 0$ (which was proved to hold at the bigger steady state) is equivalent to

$$(1-2\lambda M)(1-e^{-\beta(M+\gamma_i M_i)}) + \beta M(1-\lambda M)e^{-\beta(M+\gamma_i M_i)} < \frac{1}{\mathcal{N}}.$$

Using as before the fact that $M$ is a steady state allows us to rewrite this last inequality as

$$(1-2\lambda M)\frac{1}{\mathcal{N}}\frac{M+\gamma_i M_i}{M(1-\lambda M)} + \beta M(1-\lambda M)\big(1-\frac{M+\gamma_i M_i}{\mathcal{N}(1-\lambda M)M}\big) < \frac{1}{\mathcal{N}}.$$

Multiplying this inequality by $\mathcal{N}(1-\lambda M)\frac{M}{M+\gamma_i M_i}$ yields

$$(1-2\lambda M) + \beta(1-\lambda M)\big(\mathcal{N}M^2\frac{1-\lambda M}{M+\gamma_i M_i} - M\big) < (1-\lambda M)\frac{M}{M+\gamma_i M_i},$$

that is

$$(1-\lambda M)\Big(2-\frac{M}{M+\gamma_i M_i} + \beta M\big(-1+\mathcal{N}M\frac{1-\lambda M}{M+\gamma_i M_i}\big)\Big) < 1,$$

whence the stability of the bigger steady state, since we recover (9.20). Likewise, at the smaller steady state we have $\partial_x f > 0$, and the reverse inequality holds. This concludes the proof.

## 9.B Basin entrance time approximation

### 9.B.1 Bounds on the wild equilibria

For $M_i = 0$, under the assumptions of Lemma 9.3 such that there are two positive steady states $\mathbf{E}_- \ll \mathbf{E}_+$ for (9.2), we get explicit bounds on these states. In particular, we assume $\mathcal{N} > 4\psi$. We recall that the positive equilibria can be expressed as an increasing function of their second coordinate $M \in (0, 1/\lambda)$:

$$\mathbf{E}(M) := \begin{pmatrix} K\lambda M \\ M \\ \frac{\nu_E+\mu_E}{b}\frac{\lambda M}{(1-\lambda M)} \end{pmatrix},$$

and $\mathbf{E}(M)$ is an equilibrium if and only if $f(\beta M) = 0$, where

$$f(x) = (1-\psi x)(1-e^{-x}) - \frac{1}{\mathcal{N}}. \tag{9.21}$$

**Lemma 9.7.** *The function $f$ (defined in (9.21)) is concave on $[0, 1/\psi]$. It reaches its maximum value on this interval at $Z(\psi) \in (0, \frac{1}{2\psi})$, where we define*

$$e^{-Z(\psi)} = \frac{\psi}{1+\psi-\psi Z(\psi)}, \quad F(\psi) := \frac{1+\psi-\psi Z(\psi)}{(1-\psi Z(\psi))^2}. \tag{9.22}$$

*Then $f$ on $[0, 1/\psi]$ has no zero if $\mathcal{N} < F(\psi)$, exactly 1 zero if $\mathcal{N} = F(\psi)$ and exactly 2 zeros if $\mathcal{N} > F(\psi)$.*

*In addition, $Z$ and $F$ have the following asymptotics:*

$$Z(\psi) \sim_{\psi\to+\infty} \frac{1}{2\psi}, \quad Z(\psi) \sim_{\psi\to 0} \log\big(\frac{1}{\psi}\big), \quad F(\psi) \sim_{\psi\to+\infty} 4\psi, \quad F \xrightarrow[\psi\to 0]{} 1.$$

*Proof.* We compute

$$f'(x) = e^{-x}\big(1+\psi-\psi x\big) - \psi, \quad f''(x) = e^{-x}\big(\psi x - 1 - 2\psi\big),$$

hence $f'' < 0$ on $[0, 1/\psi]$. Since $f(0) = f(1/\psi) = -1/\mathcal{N} < 0$, $f$ reaches a unique maximum at the (necessarily unique) point $Z(\psi) \in (0, 1/\psi)$ such that $f'(Z(\psi)) = 0$. The claim that $Z(\psi) < 1/(2\psi)$ follows from the inequality $e^x > 1 + x$, which implies that

$$\frac{1}{\psi} f'\big(\frac{1}{2\psi}\big) = e^{-1/2\psi}\big(1+\frac{1}{2\psi}\big) - 1 < 0.$$

Moreover, the sign of $f(Z(\psi))$ is exactly that of $\mathcal{N} - F(\psi)$. The equivalents and limit follow from straightforward computations. □

**Remark 9.10.** *We notice that $Z$ is related to a well-known special function: let us introduce the (principal branch of the) special Lambert $W$ function, that is:*

$$W(y) = z,\ z \geq -1 \iff ze^z = y.$$

*Since if $y > 1$ then $z > 0$, we obtain*

$$Z(\psi) = \log\big(W(e^{1+1/\psi})\big).$$

Assume $\mathcal{N} > F(\psi)$ (defined in (9.22)), and denote by $x_- < x_+$ the two positive zeros of $f$.

**Lemma 9.8.** *We have $x_- > 1/\mathcal{N}$.*

$$\frac{1}{\mathcal{N}} < x_- < \frac{1}{\psi}\big(1-\frac{\kappa^*}{\mathcal{N}}\big) < Z(\psi) < \frac{1}{\psi}\big(1-\frac{\kappa_*}{\mathcal{N}}\big) < x_+,$$

*where*

$$\kappa_* = 1 + \frac{\psi}{1-\psi Z(\psi)}, \quad \kappa^* = \mathcal{N} - \frac{\psi Z(\psi)(1+\psi-\psi Z(\psi))}{(1-\psi Z(\psi))^2}.$$

*If in addition $\mathcal{N} > 2$ then $x_+ < \frac{1}{\psi}\big(1-\frac{1}{\mathcal{N}}\big)$.*

*Proof.* The first inequality is obtained by using the inequalities $1-e^{-x} \le x$ and $1-\sqrt{1-x} > x/2$ for $x \in (0,1)$. The first one implies that $f(x) \le x(1-\psi x) - 1/\mathcal{N}$, which is a second order polynomial equal to $f$ at 0 and at $1/\psi$, with roots located at $\big(1 \pm \sqrt{1-4\psi/\mathcal{N}}\big)/(2\psi)$ (recall that we have $\mathcal{N} > 4\psi$). Hence $x_- > \big(1-\sqrt{1-4\psi/\mathcal{N}}\big)/(2\psi) > 1/\mathcal{N}$ by the second inequality.

The upper bound on $x_+$ comes from the fact that if $\mathcal{N} > 2$ then by Lemma 9.7

$$\big(1-\frac{1}{\mathcal{N}}\big)\frac{1}{\psi} > \frac{1}{2\psi} > Z(\psi).$$

Finally to get the two other bounds, we introduce

$$H(\kappa) := f\big(\frac{1}{\psi}(1-\frac{\kappa}{\mathcal{N}})\big) = \kappa\big(1-e^{-\frac{1}{\psi}(1-\frac{\kappa}{\mathcal{N}})}\big) - 1.$$

By Lemma 9.7, it is concave on $[0,\mathcal{N}]$, equal to $-1$ at 0 and $\mathcal{N}$ and reaches its maximum at $\widehat{\kappa} := \mathcal{N}(1-\psi Z(\psi))$. To get $\kappa_*$ and $\kappa^*$, we simply use the fact that the graph of $H$ is above the segments from $(0,-1)$ to $(\widehat{\kappa}, H(\widehat{\kappa}))$ on the first hand, and from $(\widehat{\kappa}, H(\widehat{\kappa}))$ to $(\mathcal{N},0)$ on the other hand, so that we define

$$-1 + \frac{H(\widehat{\kappa})+1}{\widehat{\kappa}}\kappa_* = 0 = -1 - (\kappa^* - \mathcal{N})\frac{H(\widehat{\kappa})+1}{\mathcal{N}-\widehat{\kappa}},$$

and the expressions of $\kappa_* < \widehat{\kappa} < \kappa^*$ follow from a straightforward computation.

□

Back to the steady states of (9.2), we deduce from Lemma 9.8 the following bounds, assuming $\mathcal{N} > 2$:

$$\underline{\widehat{\mathbf{E}}}_- := \begin{pmatrix} \frac{\lambda K}{\mathcal{N}\beta} \\ \frac{1}{\mathcal{N}\beta} \\ \frac{\nu_E+\mu_E}{b}\frac{\lambda K}{\mathcal{N}\beta} \end{pmatrix} \le \mathbf{E}_- \le \big(1-\frac{\kappa^*}{\mathcal{N}}\big)\begin{pmatrix} K \\ \frac{1}{\lambda} \\ \frac{\nu_E+\mu_E}{b}\frac{K\mathcal{N}}{\kappa^*} \end{pmatrix} =: \widehat{\overline{\mathbf{E}}}_- \tag{9.23}$$

$$\underline{\widehat{\mathbf{E}}}_+ := \big(1-\frac{\kappa_*}{\mathcal{N}}\big)\begin{pmatrix} K \\ \frac{1}{\lambda} \\ \frac{\nu_E+\mu_E}{b}\frac{K\mathcal{N}}{\kappa_*} \end{pmatrix} \le \mathbf{E}_+ \le \big(1-\frac{1}{\mathcal{N}}\big)\begin{pmatrix} K \\ \frac{1}{\lambda} \\ \frac{K\mathcal{N}(\nu_E+\mu_E)}{b} \end{pmatrix} =: \widehat{\overline{\mathbf{E}}}_+. \tag{9.24}$$

### 9.B.2 Results

**A lower bound.** First, we give a lower bound on the entrance times. We consider the fact that for a solution to (9.2) with initial data given by $\mathbf{E}_+$, thanks to the overestimation in (9.24),

$$F(t) \ge \widehat{\underline{F}}_+ e^{-\mu_F t} =: \widehat{\underline{F}}_\flat(t).$$

This implies

$$E(t) \ge e^{-(\nu_E+\mu_E)t - \frac{b\widehat{\underline{F}}_+}{K}(1-e^{-\mu_F t})}\widehat{\underline{E}}_+ + b\widehat{\underline{F}}_+ \int_0^t e^{-\mu_F t'} e^{-(\nu_E+\mu_E)(t-t')} e^{-\frac{b\widehat{\underline{F}}_+}{K}(e^{-\mu_F t'}-e^{-\mu_F t})} dt' =: \widehat{\underline{E}}_\flat(t),$$

and

$$M(t) \ge e^{-\mu_M t}\widehat{\underline{M}}_+ + (1-r)\nu_E \int_0^t e^{-\mu_M((t-t')}\widehat{\underline{E}}_\flat(t')dt' =: \widehat{\underline{M}}_\flat(t).$$

Using the underestimation of $\mathbf{E}_-$ from (9.23), we define $t_\flat^Z := \min\{t \ge 0, \quad \widehat{\underline{Z}}_\flat(t) \le \widehat{\overline{Z}}_-\}$ for $Z \in \{E, M, F\}$.

**Lemma 9.9.** *We have the following lower bound:* $\tau(M_i) \ge \min\big(t_\flat^E, t_\flat^M, t_\flat^F\big)$.

Explicitly we find, with $Z = Z(\psi)$ and $Z_0 = 1+\psi-\psi Z$:

$$t_\flat^F = \frac{1}{\mu_F}\log\big(\frac{\kappa^*(\mathcal{N}-\kappa_*)}{\kappa_*(\mathcal{N}-\kappa^*)}\big) = \frac{1}{\mu_F}\log\big(1 + \frac{\mathcal{N}^2(1-\psi Z)^3}{\psi Z Z_0^2} - \frac{\mathcal{N}(1-\psi Z)}{\psi Z Z_0}\big).$$

However it must be expected that $\min(t_\flat^E, t_\flat^M) > t_\flat^F$, and we can give explicit approximations of $t_\flat^E$ and $t_\flat^M$.

**A first upper bound.** We compare the solution of (9.2) with the solution of the linear system

$$\begin{cases} \dfrac{dE_e}{dt} = bF_e - (\nu_E + \mu_E)E_e, \\ \dfrac{dM_e}{dt} = (1-r)\nu_E E_e - \mu_M M_e, \\ \dfrac{dF_e}{dt} = r\nu_E \epsilon(M_i) E_e - \mu_F F_e, \end{cases} \tag{9.25}$$

where $\epsilon(M_i) = \max_{t\geq 0} \frac{M(t)}{M(t)+M_i} < 1$, typically $\epsilon(M_i) = \frac{M^*}{M^*+M_i}$. The following property follows from the fact that (9.2) is cooperative:

**Lemma 9.10.** *Solutions of (9.2) and (9.25) with initial data such that $(E^0, M^0, F^0) \leq (E_e^0, M_e^0, F_e^0)$ satisfy:*

$$\forall t \geq 0,\ (E(t), M(t), F(t)) \leq (E_e(t), M_e(t), F_e(t)).$$

We use the under-estimation of $\mathbf{E}_-$ given by (9.23), to define, for $X = (X^i)_i = (E, M, F)$ and $i \in \{1, 2, 3\}$,

$$t_{\min}^{X^i} := \inf\{t \geq 0, X_e^i(t) \leq [\widehat{\underline{\mathbf{E}}}_-]_i\}.$$

**Lemma 9.11.** *For any solution $X_e$ to (9.25) satisfying the assumption of Lemma 9.10, we have the upper bound on the entrance time: $\tau(M_i) \leq \max\left(t_{\min}^E, t_{\min}^M, t_{\min}^F\right)$.*

Analytic computations are made in Section 9.B.3.

**An second upper bound in two steps.** Let $\rho^* := M_i/\widehat{\overline{M}}_+$ be the under-estimated effort ratio. When using the above one-step approach, we conclude with a finite upper bound for $\tau(M_i)$ if and only if $\widehat{\overline{M}}_+/(M_i + \widehat{\overline{M}}_+) < 1/\mathcal{N}$, that is

$$\rho^* > \mathcal{N} - 1. \tag{9.26}$$

Expanding upon the same idea as for the lower bound, we let $\epsilon = \widehat{\overline{M}}_+/(\widehat{\overline{M}}_+ + M_i)$ so

$$F(t) \leq \widehat{\overline{F}}_+ e^{-\mu_F t} + \widehat{\overline{E}}_+ r\nu_E \epsilon (1 - e^{-\mu_F t}) =: \widehat{\overline{F}}_\sharp.$$

Then, we construct the explicit solution $(E, M) = (\widehat{\overline{E}}_\sharp, \widehat{\overline{M}}_\sharp)$ to

$$\dot{E} = b\widehat{\overline{F}}_\sharp - \left(\nu_E + \mu_E + \tfrac{\widehat{\overline{F}}_\sharp}{K}\right)E, \quad E(0) = \widehat{\overline{E}}_+,$$

$$\dot{M} = (1-r)\nu_E E - \mu_M M, \quad M(0) = \widehat{\overline{M}}_+.$$

In details:

$$\widehat{\overline{F}}_\sharp(t) = \widehat{\overline{E}}_+ r\nu_E\epsilon + e^{-\mu_F t}\left(\widehat{\overline{F}}_+ - r\nu_E\epsilon\widehat{\overline{E}}_+\right),$$

$$\widehat{\overline{E}}_\sharp(t) = e^{-(\nu_E+\mu_E+\frac{\widehat{\overline{E}}_+ r\nu_E\epsilon}{K})t - \frac{\widehat{\overline{F}}_+ - r\nu_E\epsilon\widehat{\overline{E}}_+}{K\mu_F}(1-e^{-\mu_F t})}\Big(\widehat{\overline{E}}_+ + \int_0^t \big(b\widehat{\overline{E}}_+ r\nu_E\epsilon$$

$$+be^{-\mu_F t'}(\widehat{\overline{F}}_+ - r\nu_E\epsilon\widehat{\overline{E}}_+)\big)e^{(\nu_E+\mu_E+\frac{\widehat{\overline{E}}_+ r\nu_E\epsilon}{K})t' - \frac{\widehat{\overline{F}}_+ + r\nu_E\epsilon\widehat{\overline{E}}_+}{K\mu_F}(1-e^{-\mu_F t'})}dt'\Big),$$

$$\widehat{\overline{M}}_\sharp(t) = e^{-\mu_F t}\widehat{\overline{M}}_+ + (1-r)\nu_E \int_0^t e^{\mu_F t'}\widehat{\overline{E}}_\sharp(t')dt'.$$

We use this super-solution on $[0, t_0]$ (for some $t_0 > 0$ to be determined), and then glue the solution on $[t_0, +\infty)$ of

$$\begin{cases} \dot{E} = bF - (\nu_E + \mu_E)E, \quad E(t_0) = \widehat{\overline{E}}_\sharp(t_0), \\ \dot{M} = (1-r)\nu_E E - \mu_M M, \quad M(t_0) = \widehat{\overline{M}}_\sharp(t_0), \\ \dot{F} = r\nu_E\epsilon_0 E - \mu_F F, \quad F(t_0) = \widehat{\overline{F}}_\sharp(t_0), \end{cases}$$

with $\epsilon_0 = \widehat{\overline{M}}_\sharp(t_0)/(\widehat{\overline{M}}_\sharp(t_0) + M_i) < \epsilon$.

For $Z \in \{E, M, F\}$ we let

$$t_\sharp^Z(t_0) := \min\{t \geq t_0, \widehat{\overline{Z}}_\sharp \leq \widehat{\underline{Z}}_-\}.$$

Then as before:

**Lemma 9.12.** *For all $t_0 > 0$, $\tau(M_i) \leq t_\sharp(t_0) := \max(t_\sharp^E(t_0), t_\sharp^M(t_0), t_\sharp^F(t_0))$.*

By using Lemma 9.12, we can theoretically obtain a finite upper bound for $\tau(M_i)$ (upon choosing a suitable $t_0$) as soon as $\epsilon_0 < 1/\mathcal{N}$ for $t_0$ large enough, that is if and only if

$$\rho^*\big((\rho^*+1)\frac{\mu_M}{(1-r)\nu_E} + \mathcal{N} - 1\big) > \mathcal{N} - 1. \tag{9.27}$$

Condition (9.27) is weaker than (9.26) (and in general, much weaker). It holds if and only if

$$\rho^* > \frac{-(\mathcal{N}-1+\phi) + \sqrt{(\mathcal{N}-1+\phi)^2 + 4\phi(\mathcal{N}-1)}}{2\phi}, \quad \phi := \lambda K = \frac{\mu_M}{(1-r)\nu_E},$$

which is true for instance if $\rho^* > \sqrt{(\mathcal{N}-1)/\phi}$. However, we do not develop any further these analytic computations in the present paper.

### 9.B.3 Analytic computations

Applying Lemma 9.11, in order to express analytically the solution $X_e := (E_e, M_e, F_e)$ of (9.25), we only need to diagonalize the matrix

$$R_e := \begin{pmatrix} -(\nu_E + \mu_E) & b \\ r\nu_E\epsilon & -\mu_F \end{pmatrix}.$$

$R_e$ has negative trace, and positive determinant if and only if $\frac{1}{\mathcal{N}} > \epsilon$. Hence if $\mathcal{N}\epsilon(M_i) < 1$ then $\mathbf{0}$ is globally asymptotically stable for (9.25).

In this case its eigenvalues are real, negative and equal to $\kappa_\pm$ associated respectively with eigenvectors $\begin{pmatrix} 1 \\ x_\pm \end{pmatrix}$, where

$$\kappa_\pm := \frac{-(\nu_E + \mu_E + \mu_F) \pm \sqrt{(\nu_E + \mu_E - \mu_F)^2 + 4br\nu_E\epsilon}}{2},$$
$$x_\pm := \frac{\nu_E + \mu_E - \mu_F \pm \sqrt{(\nu_E + \mu_E - \mu_F)^2 + 4br\nu_E\epsilon}}{2b}.$$

Then we deduce that for some real numbers $(r_\pm^0, s_\pm^0) \in \mathbb{R}^4$,

$$E_e(t) = r_+^0 e^{\kappa_+ t} + r_-^0 e^{\kappa_- t},$$
$$F_e(t) = s_+^0 e^{\kappa_+ t} + s_-^0 e^{\kappa_- t},$$
$$M_e(t) = e^{-\mu_M t} M_e^0 + (1-r)\nu_E \int_0^t e^{-\mu_M(t-t')}\big(r_+^0 e^{\kappa_+ t'} + r_-^0 e^{\kappa_- t'}\big)dt'.$$

In details, we find

$$r_+^0 = \frac{x_-}{x_- - x_+}E_e^0 - \frac{1}{x_- - x_+}F_e^0, \qquad r_-^0 = \frac{-x_+}{x_- - x_+}E_e^0 + \frac{1}{x_- - x_+}F_e^0$$
$$s_+^0 = \frac{x_+x_-}{x_- - x_+}E_e^0 - \frac{x_+}{x_- - x_+}F_e^0, \qquad s_-^0 = \frac{-x_+x_-}{x_- - x_+}E_e^0 + \frac{x_-}{x_- - x_+}F_e^0.$$

Assuming $\kappa_+ \neq -\mu_M$ and $\kappa_- \neq -\mu_M$ (which must hold generically since these are biological parameters), we get

$$M_e(t) = e^{-\mu_M t} M_e^0 + (1-r)\nu_E\big(r_+^0 \frac{e^{\kappa_+ t} - e^{-\mu_M t}}{\mu_M + \kappa_+} + r_-^0 \frac{e^{\kappa_- t} - e^{-\mu_M t}}{\mu_M + \kappa_-}\big).$$

Assuming $\mathcal{N} > 2$, we use the overestimation (9.24) of $\mathbf{E}_+$ as an initial data $(E_e^0, M_e^0, F_e^0)$, and with the notations

$$g(\epsilon) = \sqrt{1 + \frac{4br\nu_E\epsilon}{(\nu_E + \mu_E - \mu_F)^2}}, \quad \sigma = sgn(\nu_E + \mu_E - \mu_F),$$

we deduce

$$r_\pm^0 = \frac{K}{2}\big(1 - \frac{1}{\mathcal{N}}\big)\Big(1 \pm \frac{(2\mathcal{N}-1)(\nu_E+\mu_E)+\mu_F}{g(\epsilon)|\nu_E+\mu_E-\mu_F|}\Big),$$
$$s_\pm^0 = \frac{K|\nu_E+\mu_E-\mu_F|}{4bg(\epsilon)}\big(1 - \frac{1}{\mathcal{N}}\big)\big(\sigma \pm g(\epsilon)\big)\big(g(\epsilon) \pm \frac{(2\mathcal{N}-1)(\nu_E+\mu_E)+\mu_F}{|\nu_E+\mu_E-\mu_F|}\big).$$

If $r_-^0 < 0$ the we can use the simple upper bound $E_e(t) \le r_+^0 e^{\kappa_+ t}$. This condition reads

$$g(\epsilon)|\nu_E + \mu_E - \mu_F| < (2\mathcal{N}-1)(\nu_E+\mu_E) + \mu_F.$$

In this case, we know that $E_e(t) \le \big[\widehat{\underline{\mathbf{E}}}_-\big]_1$ if $r_+^0 e^{\kappa_+ t} \le \frac{\lambda K}{\mathcal{N}\beta}$, that is if

$$t \ge t_{\min}^E := \frac{2}{\nu_E+\mu_E+\mu_F - g(\epsilon)|\nu_E+\mu_E-\mu_F|}\log\Big(\frac{(\mathcal{N}-1)}{2\psi}\big(1 + \frac{(2\mathcal{N}-1)(\nu_E+\mu_E)+\mu_F}{g(\epsilon)|\nu_E+\mu_E-\mu_F|}\big)\Big) \tag{9.28}$$

Then, under the same condition we have $s_\pm^0 > 0$. By using the fact that $s_+^0 + s_-^0 = \widehat{\overline{F}}_+$, we deduce that $F_e(t) \le \big[\widehat{\underline{\mathbf{E}}}_-\big]_3$ if $\widehat{\overline{F}}_+ e^{\kappa_+ t} \le \widehat{\underline{F}}_-$, that is if

$$t \ge t_{\min}^F := \frac{2}{\nu_E+\mu_E+\mu_F - g(\epsilon)|\nu_E+\mu_E-\mu_F|}\log\Big(\frac{\mathcal{N}(\mathcal{N}-1)}{\psi}\Big). \tag{9.29}$$

In addition, we have $t_{\min}^E > t_{\min}^F$ if and only if

$$(2\mathcal{N}-1)(\nu_E+\mu_E) + \mu_F > (\mathcal{N}-1)g(\epsilon)|\nu_E+\mu_E-\mu_F|.$$

**Remark 9.11.** *For small $\epsilon$, the previous estimations roughly show that*

$$t_{\min} \ge \frac{1}{\min(\nu_E+\mu_E, \mu_F)}\log\big(\frac{\mathcal{N}^2}{\psi}\big).$$

Finally, we need to compute the condition $M_e(t) \le \frac{1}{\mathcal{N}\beta}$. Let $\sigma_E := \mu_M/(\nu_E+\mu_E)$ and $\sigma_F := \mu_M/\mu_F$. We rewrite $M_e(t)$ as

$$M_e(t) = \frac{1}{\lambda}\big(1 - \frac{1}{\mathcal{N}}\big)\big(\alpha e^{-\mu_M t} + \alpha_+ e^{\kappa_+ t} + \alpha_- e^{\kappa_- t}\big),$$

with

$$\alpha = \frac{(\mathcal{N}-1)\sigma_F + 1 - \epsilon\mathcal{N}}{(\sigma_F-1)(\sigma_E-1) - \epsilon\mathcal{N}}, \quad \alpha_\pm = \frac{\mu_M}{\mu_M+\kappa_\pm}\widetilde{r}_\pm^0,$$

where

$$\widetilde{r}_\pm^0 := \frac{1}{2}\big(1 \pm \frac{2\mathcal{N}-1+\sigma_E/\sigma_F}{g(\epsilon)\sigma(1-\sigma_E/\sigma_F)}\big), \quad g(\epsilon) = \sqrt{1 + \frac{4\mathcal{N}\sigma_E\sigma_F\epsilon}{(\sigma_F-\sigma_E)^2}}$$

and

$$\frac{\mu_M}{\mu_M+\kappa_\pm} = \frac{2\sigma_E\sigma_F}{2\sigma_E\sigma_F - (\sigma_E+\sigma_F) \pm \sigma(\sigma_F-\sigma_E)g(\epsilon)}.$$

The condition we need to compute is therefore

$$\alpha e^{-\mu_M t} + \alpha_+ e^{\kappa_+ t} + \alpha_- e^{\kappa_- t} \le \frac{\psi}{\mathcal{N}-1}.$$

We assume that the male half-life is shorter than that of the females and of the eggs, so that $\sigma_F, \sigma_E > 1$. Under the stronger assumptions that $r_- < 0 < r_+$ and

$$\epsilon\mathcal{N} < 1, \quad (\sigma_F-1)(\sigma_E-1) > \epsilon\mathcal{N},$$

we obtain that $\alpha > 0$. We simply treat two subcases: first if $\mu_M + \kappa_+ < 0$ (small $\mu_M$) then we obtain $\alpha_+ < 0 < \alpha_-$ and thus

$$t^M_{\min} := \frac{1}{\mu_M} \log\Big((\mathcal{N}-1)\frac{\alpha+\alpha_-}{\psi}\Big).$$

Second, if $\mu_M + \kappa_- > 0$ (large $\mu_M$) then we obtain $\alpha_- < 0 < \alpha_+$ and thus

$$t^M_{\min} := \frac{1}{-\kappa_+} \log\Big((\mathcal{N}-1)\frac{\alpha+\alpha_+}{\psi}\Big).$$

In the last case (when $\mu_M$ is large), we can check that $t^M_{\min} > t^E_{\min}$ is equivalent to

$$\alpha + \alpha_+ > \widetilde{r}^0_+,$$

which holds since $\alpha > 0$ and $\alpha_+ > \widetilde{r}^0_+$.

In this case we obtain

$$\begin{aligned}\max\big(t^E_{\min}, t^F_{\min}, t^M_{\min}\big) = t^M_{\min}& \\ = \frac{2\sigma_E}{\mu_F\big(\sigma_F+\sigma_E - g(\epsilon)\sigma(\sigma_F-\sigma_E)\big)} \log\Big(\frac{\mathcal{N}-1}{\psi}\big(&\frac{(\mathcal{N}-1)\sigma_F + 1 - \epsilon\mathcal{N}}{(\sigma_F-1)(\sigma_E-1)-\epsilon\mathcal{N}} \\ &+ \frac{\sigma_E\sigma_F\big(g(\epsilon)\sigma(\sigma_F-\sigma_E) + (2\mathcal{N}-1)\sigma_F + \sigma_E\big)}{\big(2\sigma_E\sigma_F - (\sigma_E+\sigma_F) + \sigma(\sigma_F-\sigma_E)g(\epsilon)\big)g(\epsilon)\sigma(\sigma_F-\sigma_E)}\big)\Big).\end{aligned}$$

# Chapter 10

# Optimal releases for population replacement strategies, application to *Wolbachia*

> 'So the problem will solve itself,' said Philip.
> 'Only by destroying itself. When humanity's destroyed, obviously there'll be no more problem. But it seems a poor sort of solution. I believe there may be another, even within the framewok of the present system. A temporary one while the system's being modified in the direction of a permanent solution. (...)'
>
> Aldous Huxley, *Point Counter Point.*

This chapter is a joint work with Luis Almeida, Yannick Privat and Nicolas Vauchelet.

**Abstract.** In this article, we consider a simplified model of time dynamics for a mosquito population subject to the artificial introduction of *Wolbachia*-infected mosquitoes, in order to fight arboviruses transmission. Indeed, it has been observed that when some mosquito populations are infected by some *Wolbachia* bacteria, various reproductive alterations are induced in mosquitoes, including cytoplasmic incompatibility. Some of these *Wolbachia* bacteria greatly reduce the ability of insects to become infected with viruses such as the dengue ones, cutting down their vector competence and thus effectively stopping local dengue transmission.

The behavior of infected and uninfected mosquitoes is assumed to be driven by a compartmental system enriched with the presence of an internal control source term standing for releases of infected mosquitoes, distributed in time. We model and design an optimal releasing control strategy with the help of a least square problem. In a nutshell, one wants to minimize the number of uninfected mosquitoes at a given horizon of time, under some relevant biological constraints. We derive properties of optimal controls, highlight a limit problem providing useful asymptotic properties of optimal controls. We numerically illustrate the relevance of our approach.

## 10.1 Introduction

For many years (since [110]), scientists have been studying *Wolbachia*, a bacterium living only inside insect cells. Recently, there has been increasing interest in the biology of *Wolbachia* and in its application as an agent for control of vector mosquito populations, by taking advantage of a phenomenon called *cytoplasmic incompatibility*. In key vector species such as *Aedes aegypti*, if a male mosquito infected with *Wolbachia* mates with a non-infected female, the embryos die early in development, in the first mitotic divisions (see [233]). This also happens even if the male and female are both infected with *Wolbachia* but are carrying mutually incompatible strains. Interestingly, an infected female can mate with an uninfected male producing healthy eggs just fine. Hence, using *cytoplasmic incompatibility* (CI) allows scientists to produce functionally sterile males that can be released in the field as an elimination tool against mosquitoes. This vector control method is known as incompatible insect technique (IIT).

Another promising application of this symbiotic bacteria is the control of endemic mosquito-borne diseases by means of population replacement. This control relies on the *pathogen interference* (PI) phenotype of some *Wolbachia* strains, especially with Zika, dengue and chikungunya viruses in *Aedes* mosquitoes (see [232]). Population replacement methods have the benefit of being more environmentally benign than insecticide-based approaches (since they are species specific) and potentially more cost effective (since they are long-lasting). Despite the broad range of arthropods carrying *Wolbachia*, no transmission event to any warm-blooded animals has been reported. The principle is to release *Wolbachia* carrying mosquitoes in endemic areas. Once released, they breed with wild mosquitoes. Over time and if releases are large and long enough, one can expect the majority of mosquitoes to carry *Wolbachia*, thanks to CI. Due to PI, the mosquito population then has a reduced vector competence, decreasing the risk of Zika, dengue and chikungunya outbreaks.

Both IIT and population replacement procedure have been imagined since a long time (see e.g. the work by Laven in 1967 [143] for population replacement, or the one by Curtis and Adak [61] in 1974 for population elimination, both on mosquitoes in genus *Culex*), but there has been a resurgence of interest lately for both techniques due to the increasing burden of arboviral diseases transmitted by mosquitoes in genus *Aedes*, and their operational implementation is a hot topic since the first report in [118] of field success in Australian *Aedes aegypti* (see [144] for IIT). We focus here on population replacement strategies.

Motivated by the issue of controlling a population of wild *Aedes* mosquitoes by means of *Wolbachia* infected ones, we investigate here a simplified control model of population replacement strategies, where one acts on the wild population by means of time-distributed releases of infected individuals. The evolution equations we use incorporates the competition of released individuals with the wild ones. Formally, let $n_1(t)$ denote the density of *Wolbachia*-free mosquitoes (the wild individuals) and $n_2(t)$ the density of *Wolbachia*-infected mosquitoes (the introduced ones) at time $t$. We model population densities dynamics by the following competitive compartmental system:

$$\begin{cases} \dfrac{dn_1}{dt}(t) = f_1(n_1(t), n_2(t)), \\ \dfrac{dn_2}{dt}(t) = f_2(n_1(t), n_2(t)) + u(t), \qquad t > 0, \\ n_1(0) = n_1^0, \quad n_2(0) = n_2^0, \end{cases} \tag{10.1}$$

where $u(\,\cdot\,)$ is a non-negative function standing for a **control** (it models the release of *Wolbachia*-infected mosquitoes). The terms $f_i(n_1, n_2)$, $i = 1, 2$, are defined by

$$f_1(n_1, n_2) = b_1 n_1 \big(1 - s_h \frac{n_2}{n_1 + n_2}\big)\big(1 - \frac{n_1 + n_2}{K}\big) - d_1 n_1, \tag{10.2}$$

$$f_2(n_1, n_2) = b_2 n_2 \big(1 - \frac{n_1 + n_2}{K}\big) - d_2 n_2. \tag{10.3}$$

The term $(1 - s_h \frac{n_2}{n_1+n_2})$ models the cytoplasmic incompatibility (CI): the parameter $s_h$ is the CI rate; one has $0 \leq s_h \leq 1$ and when $s_h = 1$, CI is perfect, whereas when $s_h = 0$ there is no CI. The other parameters $(b_i, d_i)$ for $i \in \{1, 2\}$ are respectively mortality and birth rates, and $K$ denotes the environmental carrying capacity.

A model such as (10.2)-(10.3) for mosquito population dynamics with *Wolbachia* has been introduced in [83], and also studied [121] where it was coupled with an epidemiological model. In [50], similar dynamics have been described (including also a spatial dimension); further discussion on these various models can be found in [211]. We note that the addition of a control term was already proposed in [44] for population replacement and in [216] for IIT (coupled with insecticide), where some associated optimization problems were described.

To make it closed, this system is complemented with nonnegative initial data $(n_1^0, n_2^0)$ and we will assume to be, at time $t = 0$, in the "worst" initial situation where there are no *Wolbachia*-infected mosquitoes in the population, in other words $n_2^0 = 0$. When useful, we will use the notations

$$\mathbf{n} = (n_1, n_2) \qquad \text{and} \qquad \mathbf{f} = (f_1, f_2)$$

to denote respectively the density mosquitoes vector and the right-hand side functions vector in (10.2)-(10.3).

The mathematical model (10.1)-(10.2)-(10.3) in the absence of control (in other words when $u = 0$) will be analyzed and commented in Section 10.2.1. The starting point of our analysis is to

notice that this system has, as steady states (in addition to the trivial one $(0,0)$)

$$(n_1^*, 0) \quad \text{and} \quad (0, n_2^*), \quad \text{with } n_i^* = K\left(1 - \frac{b_i}{d_i}\right), \ i = 1, 2,$$

corresponding to the invasion of the total population of mosquitoes, either by the wild ones or the *Wolbachia*-infected one. In the following, we will make several assumptions guaranteeing that System (10.1) is bistable and monotone. Our main objective is to build a strategy allowing us to reach the stable state $(0, n_2^*)$, starting from the other stable state $(n_1^*, 0)$, by determining *in an optimal way* a **control law** $u(t)$. Any path leading from $(n_1^*, 0)$ to the basin of attraction of $(0, n_2^*)$ will be called a *population replacement strategy*. Our aim is thus to steer the control system as closely as possible to the steady state $(0, n_2^*)$ at time $T > 0$. In an informal way, we investigate the following issue:

*How to design optimally the releases of Wolbachia-carrying mosquitoes (in other words, how to choose a good control function $u(\cdot)$) in order to favor the establishment of Wolbachia infection?*

Of course, to make this issue relevant, it is necessary to assume some constraints on the control function $u(\cdot)$, modeling in particular the fact that the ability of scientists to create *Wolbachia*-infected mosquitoes is limited. In the converse case, it is likely that a trivial answer would be to release the maximal possible number of mosquitoes at each time $t$. In the sequel, we will hence consider the following constraints (of pointwise and integral types) on the control function $u(\cdot)$

$$0 \le u(t) \le M \text{ a.e. on } (0,T) \qquad \text{and} \qquad \int_0^T u(t)\,dt \le C$$

for some positive constants $M$ and $C$, meaning that the flux of *Wolbachia*-infected mosquitoes that can be released at each time $t$ is limited, as well as their total amount over the horizon of time $T$.

In the analysis to follow, we use the essential property that System (10.1) is competitive, meaning that it enjoys a comparison principle (see Lemma 10.1).

From the mathematical point of view, problems investigated within this article are related to optimal control theory for biological systems. Such kind of application has not been much investigated at this time. We nevertheless mention [45, 140, 218] on optimal control problems for mono/bi-stable systems, on the understanding that this list is far from being exhaustive.

Let us describe our main results. When $b_1$, $b_2$ are large, we show that the proportion of *Wolbachia*-infected mosquitoes $p = n_2/(n_1 + n_2)$ converges to the solution of a reduced problem of the form

$$\frac{dp}{dt} = f(p) + ug(p), \tag{10.4}$$

with $g \ge 0$ and $f$ of bistable type[1]. Bistable frequency-based models such as (10.4) have been studied extensively (see in particular [29]) for cytoplasmic incompatibility modeling since the works of Caspari and Watson [47]. Yet, as a new feature (10.4) incorporates rigorously a control term. The typical control for this biological system being the releases of individuals, it was unclear to understand how that control would act on the proportion $p$ of infected individuals. Our approach thus provides a way to derive a relevant control system on $p$ from the standard control system (10.1) where the input is a density of released individuals. We first prove that the optimization problems converge along with the equations ($\Gamma$-convergence result stated in Proposition 10.2) to a limit problem, and then solve it completely (Theorem 10.1). It appears that the solutions to the limit problem consist of a single release phase where the maximal flux capacity $M$ is used. Generically, this phase occurs either at the very beginning or at the very end of the time frame $[0,T]$, depending on whether the constraints allow for the existence of a population replacement strategy or not.

Numerical investigations illustrate this behavior and also hint that the optimal strategies for steering system (10.1) toward infection establishment may differ significantly from those suitable for (10.4).

The article is organized as follows. Section 10.2 is devoted to modeling issues: we introduce the simplified dynamics we consider for the system of wild versus *Wolbachia*-carrying mosquitoes, as well as the optimal control problem ($\mathcal{P}_{\text{full}}$) used to design a release strategy.

[1] The wording "**bistable function**" means that $f(0) = f(1) = 0$ and there exists $\theta \in (0,1)$ such that $f(x)(x-\theta) < 0$ on $(0,1)$ (in particular, one has necessarily $f(\theta) = 0$ whenever $f$ is smooth).

This problem is then analyzed in Section 10.3. More precisely, we show in Section 10.3.1 that ($\mathcal{P}_{\text{full}}$) and its solution converge to a population replacement strategy optimization problem ($\mathcal{P}_{\text{reduced}}$) for the simplified model (10.4), in the limit when birth rates are assumed to be large. Numerical experiments validating our approach are presented in Section 10.3.2. Some additional qualitative properties of the solutions to ($\mathcal{P}_{\text{full}}$) are proved in Appendix 10.B.

For the sake of readability, all the proofs are postponed to Appendix 10.A.

## 10.2 Toward an optimal control problem

### 10.2.1 On the dynamics without control

First we describe precisely the asymptotic behavior of System (10.1) in the absence of control (in other words, when $u(\,\cdot\,) = 0$). An example of phase portrait illustrating this lemma is provided on Fig 10.1. There and for all numerical illustrations of our results, the parameter values we choose for $b_i, d_i$ and $s_h$ reflect the effects of a *Wolbachia* infection in *Aedes* mosquitoes. In the well-documented case of the *Wolbachia* strain *wMel* in *Aedes aegypti* and according to [232, 75], it is relevant to choose: slight fecundity reduction ($b_2/b_1 \simeq 0.9$), slight life-span reduction ($d_2/d_1 \simeq 1.1$) and almost perfect CI ($s_h = 0.9$). We do not fix a time scale, hence the last biologically meaningful parameter is $b_1/d_1$, the basic reproduction number for the wild population. Freely inspiring from literature estimates (see [88, 181, 121]) we assume that this number is large, at least equal to 3 (and describing all the range $[3.7, 7400]$ in Section 7.5). Since these values are used only for results illustration, they are not intended to represent precisely a well-identified mosquito population-*Wolbachia* strain couple.

**Lemma 10.1.** *System (10.1) is positive and (monotone) competitive*[2].

*Let us assume that*

$$b_1 > d_1 \quad \text{and} \quad b_2 > d_2. \tag{10.5}$$

*Then, System (10.2)-(10.3) with $u(\,\cdot\,) = 0$ has at least three non-negative steady states:*

$$(0,0), \quad (n_1^*, 0), \quad (0, n_2^*), \quad \text{with} \quad n_i^* = K\left(1 - \frac{d_i}{b_i}\right), \ i \in \{1,2\}.$$

*In this case, each population can sustain itself in the absence of the other one. In addition, $(0,0)$ is (locally linearly) unstable.*

*Moreover, there exists a fourth distinct positive steady state if and only if*

$$1 - s_h < \frac{d_1 b_2}{d_2 b_1} < 1. \tag{10.6}$$

*In this case, this coexistence equilibrium is (locally linearly) unstable, and is given by*

$$\mathbf{n}^C = K\left(\left(1 - \frac{1}{s_h}\left(1 - \frac{d_1 b_2}{d_2 b_1}\right)\right)\left(1 - \frac{d_2}{b_2}\right), \frac{1}{s_h}\left(1 - \frac{d_1 b_2}{d_2 b_1}\right)\left(1 - \frac{d_2}{b_2}\right)\right).$$

*Moreover, the two other nontrivial steady states are locally asymptotically stable in this case.*

Notice that conditions (10.5) and (10.6) on the parameters are relevant since *Wolbachia*-infected *Aedes* mosquitoes typically have (even slightly) reduced fecundity and lifespan (for instance in the case of *wMel* strain, [232]). Moreover CI is almost perfect in these species-strain combination (see [75]), *i.e.* $s_h$ is close to 1.

**Interpretation.** In short, under the biologically relevant conditions (10.5) and (10.6), the two mutual exclusion steady states are stable while whole population extinction and coexistence state are unstable: in our model, either one of the two phenotypes must prevail in the long run, eliminating the other one.

[2] This means that if $(n_1^{\pm}, n_2^{\pm})$ are solutions of (10.1) such that $n_1^-(0) < n_1^+(0)$ and $n_2^-(0) > n_2^+(0)$ then one has $n_1^-(t) < n_1^+(t)$ and $n_2^-(t) > n_2^+(t)$ for every time $t \in [0,T]$, where $(n_1^-, n_2^-)$ (resp. $(n_1^+, n_2^+)$) denotes the solution of System (10.1) associated to the choice of initial conditions $(n_1^0, n_2^0) = (n_1^-, n_2^-)$ (resp. $(n_1^0, n_2^0) = (n_1^+, n_2^+)$)

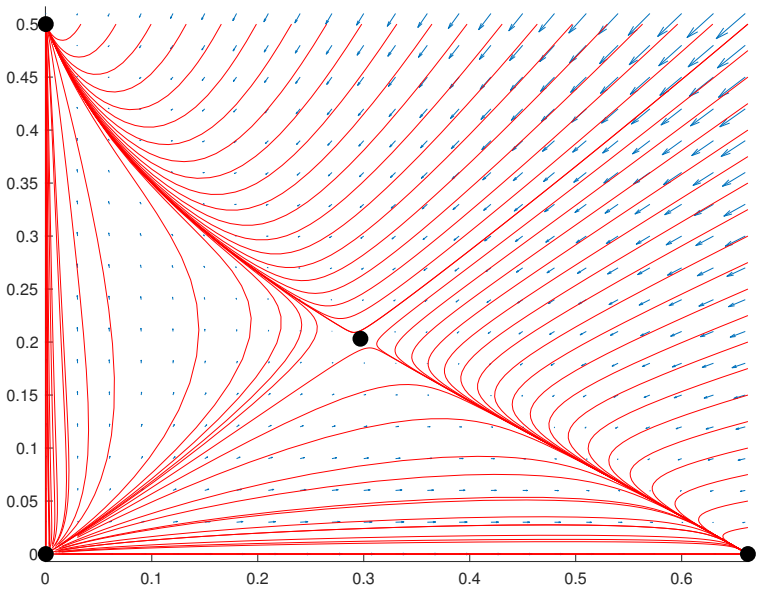



Figure 10.1: Phase portrait of System (10.1) for the parameters choice: $b_1 = 0.8$, $b_2 = 0.6$, $d_1 = 0.27$, $d_2 = 0.3$, $s_h = 0.8$ and $K = 1$ for which conditions (10.5) and (10.6) are satisfied. Examples of trajectories are plotted with continuous lines. The dots locate the four steady states.

### 10.2.2 Objective function and constraints on the control

Let us fix a horizon of time $T > 0$. In this section, we propose a relevant choice of objective function $u \mapsto J(u)$, trying to model that we expect the control be chosen so that the final state (at time $T$) of System (10.1) be as close as possible to the steady state $(0, n_2^*)$ corresponding to a population replacement situation. Since there is no obvious choice, we will consider a least square type functional, having the property to decrease as $\mathbf{n}(T)$ gets closer to $(0, n_2^*)$.

This leads to introduce

$$J(u) = \frac{1}{2} n_1(T)^2 + \frac{1}{2}(n_2^* - n_2(T))_+^2, \tag{10.7}$$

where the we used the notation $X_+ = \max\{X, 0\}$ for $X \in \mathbb{R}$ and $\mathbf{n} = (n_1, n_2)$ is the solution of (10.1) associated to, in some sense, the worst initial data $\mathbf{n}(0) = (n_1^*, 0)$. Notice that, to ensure consistency of our model, any larger value of the introduced population than the equilibrium value $n_2^*$ is beneficial for $J(u)$. This objective function differs from the ones introduced in [44], where a $L^2$ norm is used to optimize a similar protocol of *Wolbachia* infection establishment by releases. Here, we are only interested in the state at the end of the treatment, which determines protocol success or failure.

Let us enumerate the mathematical constraints we will assume on the control function $u(\cdot)$, stemming from biology.

- $u(t)$ corresponds to the density of *Wolbachia*-infected released mosquitoes and must be non-negative (since we assume that we only release individuals and cannot remove them).
- Since System (10.1) is monotone, it is relevant to assume an upper bound on the total number of released individuals, namely
$$\int_0^T u(t)dt \leq C$$
for some given $C > 0$. Indeed, releasing more and more individuals can never be detrimental. Without such a constraint, the solution of the considered optimal control problem is trivial and consists in releasing as much individuals as possible at each time.

- For practical reasons, it is neither possible to create an infinite number of *Wolbachia*-infected individuals nor to release them "instantly" at time $t$. Hence, this leads to assume a pointwise upper bound on the control, by setting $u(t) \leq M$ for some $M > 0$ and all $t \in [0,T]$. This constraint models that a release is necessarily distributed in time (possibly on a very short period of time) and cannot be an impulse.

All these considerations lead us to introduce the following set of admissible controls

$$\boxed{\mathcal{U}_{T,C,M} = \{u \in L^\infty([0,T]), \quad 0 \leq u \leq M \text{ a.e. }, \int_0^T u(t)\,dt \leq C\}}. \tag{10.8}$$

We then deal with the following optimal control problem.

$$\boxed{\inf_{u \in \mathcal{U}_{T,C,M}} J(u).} \tag{$\mathcal{P}_{\text{full}}$}$$

where $J$ is defined by (10.7) and $\mathcal{U}_{T,C,M}$ is defined by (10.8).

**Interpretation.** Problem ($\mathcal{P}_{\text{full}}$) amounts to finding a constrained release protocol (in terms of total number of released individuals and maximal release flux) which steers the system as close as possible to the target state: elimination of the wild phenotype and establishment of the introduced one.

### 10.2.3 System and problem reductions

From a practical point of view, it appears relevant to consider that birth rates are large compared with death rates, since vector *Aedes* species typically have a very high reproductive power. For this reason, we will introduce (at the end of this section) and then analyze (in Section 10.3) a simplified version of Problem ($\mathcal{P}_{\text{full}}$) that will help to infer some interesting qualitative properties of the solution of Problem ($\mathcal{P}_{\text{full}}$). This way, we will reduce System (10.1) into a simple scalar equation on the proportion of *Wolbachia*-infected mosquitoes in the spirit of [211]. To do so, let us introduce a small parameter $\epsilon > 0$ and the birth rates

$$b_1 = b_1^0/\epsilon \quad \text{and} \quad b_2 = b_2^0/\epsilon \tag{10.9}$$

for some positive numbers $b_1^0$, $b_2^0$.

It is notable that, in that case, the steady-states $(n_1^*, 0)$ and $(0, n_2^*)$ respectively converge to $(K,0)$ and $(0,K)$ as $\epsilon \searrow 0$, since $n_i^* = K(1 - \epsilon \frac{d_i}{b_i^0})$, $i = 1,2$. Notice also that (10.5) is automatically satisfied as soon as $\epsilon$ is small enough.

In what follows, we will denote by $J^\epsilon$ the functional defined by

$$J^\epsilon(u) = \frac{1}{2} n_1^\epsilon(T)^2 + \frac{1}{2}(n_2^* - n_2^\epsilon(T))_+^2, \tag{10.10}$$

where $(n_1^\epsilon, n_2^\epsilon)$ denote the solution to Problem (10.1) with $b_1$ and $b_2$ given by (10.9). Let us introduce the variables

$$N^\epsilon = n_1^\epsilon + n_2^\epsilon \qquad \text{and} \qquad p^\epsilon = n_2^\epsilon / N^\epsilon. \tag{10.11}$$

Setting $n^\epsilon = \frac{1}{\epsilon}\left(1 - \frac{N^\epsilon}{K}\right)$, we have the following (technical but crucial) convergence result, saying that the pair $(n^\epsilon, p^\epsilon)$ converges in some sense to a well-identified limit $(u,p)$.

**Proposition 10.1.** *Let $u^\epsilon \in \mathcal{U}_{T,C,M}$ such that $(u^\epsilon)_{\epsilon>0}$ converges weakly-star[3] to $u \in \mathcal{U}_{T,C,M}$ in $L^\infty(0,T)$ as $\epsilon \searrow 0$.*

*The pair $(n^\epsilon, p^\epsilon)$ associated to the control $u^\epsilon$ and the parameter scaling (10.9) solves a slow-fast system of the form*

$$\begin{cases} \epsilon \dfrac{dn^\epsilon}{dt} = (1 - \epsilon n^\epsilon) a(p^\epsilon)\big(Z(p^\epsilon) - n^\epsilon\big) - \dfrac{u^\epsilon}{K}, \\ \dfrac{dp^\epsilon}{dt} = p^\epsilon(1-p^\epsilon)\big(n^\epsilon(b_2^0 - b_1^0(1 - s_h p^\epsilon)) + d_1 - d_2\big) + \dfrac{u^\epsilon(1-p^\epsilon)}{K(1-\epsilon n^\epsilon)}, \quad t > 0 \\ n^\epsilon(0) = \dfrac{d_1^0}{b_1^0}, \; p^\epsilon(0) = 0, \end{cases} \tag{10.12}$$

[3] This means that

$$\forall v \in L^1(0,T), \qquad \int_0^T u^\epsilon v \to \int_0^T uv \quad \text{as } \epsilon \searrow 0.$$

*where $a(p)$ and $Z(p)$ are defined by*

$$a(p) = b_1^0(1-p)(1-s_h p) + b_2^0 p > 0, \quad Z(p) = \frac{d_1(1-p)+d_2 p}{a(p)} > 0.$$

*Let us assume that (10.6) holds and let $\epsilon_0 > 0$ be such that*

$$\frac{d_1}{b_1^0} < \frac{1}{\epsilon_0} \quad \text{and} \quad \max_{[0,1]} Z < \frac{1}{\epsilon_0}. \tag{10.13}$$

*Then for all $\epsilon \in (0, \epsilon_0)$ we have the uniform estimates*

$$0 \le p^\epsilon(t) \le 1 \quad \text{and} \quad n_- \le n^\epsilon(t) \le n_+ \tag{10.14}$$

*for all $t \in [0,T]$ where*

$$\begin{aligned} n_- &= \min\left\{ \frac{d_1}{b_1^0}, \min_{\epsilon\in[0,\epsilon_0]} \min_{p\in[0,1]} \frac{1+\epsilon Z(p) - \sqrt{(1-\epsilon Z(p))^2 + 4\epsilon M/(Ka(p))}}{2\epsilon} \right\} \\ n_+ &= \max\left\{ \frac{d_1}{b_1^0}, \max_{p\in[0,1]} Z(p) \right\}. \end{aligned}$$

*Then up to a subfamily, $(p^\epsilon)_{\epsilon>0}$ converges uniformly to $p$ as $\epsilon \searrow 0$, where $p$ is the solution to*

$$\begin{cases} \dfrac{dp}{dt} = p(1-p)\dfrac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} + \dfrac{u}{K}\dfrac{b_1^0(1-p)(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p}, \quad t>0 \\ p(0) = 0. \end{cases} \tag{10.15}$$

**Interpretation.** Proposition 10.1 is a rigorous result showing that a single equation on the proportion of *Wolbachia*-carrying mosquitoes (equation (10.15)) is a fair approximation of the time dynamics induced by the model with two populations (10.1), provided that the fecundity is large.

**Remark 10.1.** *It is notable that the function $[0,\epsilon_0] \ni \epsilon \mapsto \frac{1+\epsilon Z(p) - \sqrt{(1-\epsilon Z(p))^2 + 4\epsilon M/(Ka(p))}}{2\epsilon}$ used to define $n_-$ in the statement of Proposition 10.1 above converges to the finite (and bounded in $p \in [0,1]$) value $Z(p) - M/(Ka(p))$ as $\epsilon \to 0$. Therefore $n_-$ is uniformly bounded for $\epsilon \in [0,\epsilon_0]$.*

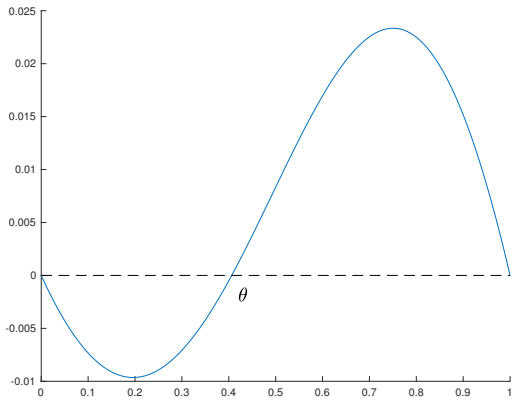



Figure 10.2: Equation (10.15) is the form $\frac{dp}{dt} = f(p) + ug(p)$, with $f$ of bistable type (see Footnote 1). Plot of the right-hand side function $f$ with the same parameters values as in Figure 10.1.

We are now in position to determine the asymptotic behavior of the solutions of Problem ($\mathcal{P}_{\text{full}}$) as $\epsilon \searrow 0$, in the case where (10.9) is assumed.

We have already observed that the invasion equilibrium $(0, n_2^*)$ is in particular changed into $(0, K(1-\epsilon\frac{d_2}{b_2^0}))$, which converges to $(0,K)$ as $\epsilon \searrow 0$. By using the result stated in Proposition 10.1,

we formally infer that $(N^\epsilon(T))_{\epsilon>0}$ converges to $K$ and $(p^\epsilon(T))_{\epsilon>0}$ converges to some limit $p(T) \in [0,1]$ as $\epsilon \searrow 0$, meaning that $(n_1^\epsilon(T), n_2^\epsilon(T)))_{\epsilon>0}$ converges to $(K(1-p(T)), Kp(T))$. It follows that $J^\epsilon(u)$ converges, as $\epsilon \searrow 0$ to

$$\frac{K^2}{2}(1-p(T))^2 + \frac{K^2}{2}(1-p(T))^2 = K^2(1-p(T))^2,$$

where $p$ denotes the solution of (10.15).

This leads to introduce the cost function (10.7) defined by

$$J^0(u) = K^2(1-p(T))^2, \tag{10.16}$$

as well as an asymptotic version of Problem ($\mathcal{P}_{\text{full}}$) reading

$$\boxed{\inf_{u\in\mathcal{U}_{T,C,M}} (1-p(T))^2,} \tag{$\mathcal{P}_{\text{reduced}}$}$$

where $p$ solves (10.15) and $\mathcal{U}_{T,C,M}$ is defined by (10.8).

In Section 10.3, we will analyze the connections between Problem ($\mathcal{P}_{\text{full}}$) and Problem ($\mathcal{P}_{\text{reduced}}$), by providing a partial description of minimizers and highlighting good convergence properties as $\epsilon \searrow 0$.

## 10.3 Analysis of Problem ($\mathcal{P}_{\text{full}}$) and numerics

### 10.3.1 Description of minimizers

This section is devoted to the analysis of Problems ($\mathcal{P}_{\text{full}}$) and ($\mathcal{P}_{\text{reduced}}$). It mainly contains two results:

- In Prop. 10.2, we state a $\Gamma$-convergence type result relating the asymptotic behavior of the solutions of Problem ($\mathcal{P}_{\text{full}}$) to the ones of Problem ($\mathcal{P}_{\text{reduced}}$). We also investigate existence issues for these problems.
- In Theorem 10.1, we completely describe the solutions of Problem ($\mathcal{P}_{\text{reduced}}$).

**Proposition 10.2.** *Let $T, C, M > 0$ and assume that (10.5) and (10.6) hold. Problem ($\mathcal{P}_{\text{full}}$) and Problem ($\mathcal{P}_{\text{reduced}}$) have (at least) a solution.*

*Moreover, let $(u^\epsilon)_{\epsilon>0}$ be a family of minimizers for Problem ($\mathcal{P}_{\text{full}}$). Then, one has*

$$\lim_{\epsilon\searrow 0} \inf_{u\in\mathcal{U}_{T,C,M}} J^\epsilon(u) = \inf_{u\in\mathcal{U}_{T,C,M}} J^0(u)$$

*and any closure point of this family (as $\epsilon \searrow 0$, for the $L^\infty$-weak star topology) is a solution of Problem ($\mathcal{P}_{\text{reduced}}$).*

**Interpretation.** Proposition 10.2 establishes that the controlled scalar equation (10.15) is not only a fair approximation of the time dynamics of the infection frequency $n_2/(n_1+n_2)$ from system (10.1), but also provides a sound framework for studying optimization problems. Morally, a release protocol defined by solving the simpler problem ($\mathcal{P}_{\text{reduced}}$) will be typically good for ($\mathcal{P}_{\text{full}}$) as well, provided that the fecundity is large.

We now solve Problem ($\mathcal{P}_{\text{reduced}}$) involving $p$, the solution to (10.4), in other words

$$\frac{dp}{dt} = f(p) + ug(p),$$

with

$$f(p) = p(1-p)\frac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} \quad \text{and} \quad g(p) = \frac{1}{K}\cdot\frac{b_1^0(1-p)(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p}. \tag{10.17}$$

In what follows, we will mainly use structural properties of $f$ and $g$, namely that they are $C^1$ functions on $[0,1]$ such that $g>0$ on $[0,1)$, $g(1)=0$, and under assumption (10.6) $f$ is a bistable

function (see Footnote 1). In what follows, we assume that (10.6) is satisfied and we denote by $\theta$ the unique real number satisfying

$$f(\theta) = 0 \qquad \text{and} \qquad \theta \in (0,1),$$

where $f$ is given by (10.17), in other words,

$$\theta = \frac{1}{s_h}\left(1 - \frac{d_1 b_2^0}{d_2 b_1^0}\right). \tag{10.18}$$

**Theorem 10.1.** *Let $T$, $C$, $M$ be three positive numbers and assume that $T > C/M$ (in other words that the horizon of time is large enough). Let us assume that (10.6) is satisfied. Any solution $u$ to ($\mathcal{P}_{\text{reduced}}$) satisfies $\int_0^T u^*(t)dt = C$ and is bang-bang (i.e. equal a.e. to 0 or $M$).*

*If $M \leq \max_{p\in[0,\theta]} -f(p)/g(p)$ then the unique solution to ($\mathcal{P}_{\text{reduced}}$) is given by $M\mathbb{1}_{[T-C/M,T]}$. Otherwise, defining*

$$C^*(M) = \int_0^\theta \frac{M dp}{f(p) + Mg(p)}, \tag{10.19}$$

*one has*

- *if $C < C^*(M)$ then the solution to ($\mathcal{P}_{\text{reduced}}$) is unique and equal to $u^* = M\mathbb{1}_{[T-C/M,T]}$. In this case $J^0(u^*) > (1-\theta)^2$;*
- *if $C > C^*(M)$ then the solution to ($\mathcal{P}_{\text{reduced}}$) is unique and equal to $u^* = M\mathbb{1}_{[0,C/M]}$. In this case $J^0(u^*) < (1-\theta)^2$;*
- *if $C = C^*(M)$ then there is a continuum of solutions to ($\mathcal{P}_{\text{reduced}}$) given by $u^*_\lambda = M\mathbb{1}_{[\lambda,\lambda+C/M]}$ for $\lambda \in [0, T - C/M]$, with $J^0(u^*_\lambda) = (1-\theta)^2$,*

*where $\theta$ is given by (10.18).*

Theorem 10.1 is illustrated on Fig. 10.3.

**Interpretation.** Theorem 10.1 implies that the best release protocol in the framework of the frequency model (10.15) consists in a single release phase, either at the beginning of the time frame if the desirable state is reachable, or at the end otherwise. To what extent must this strategy be adapted when $\epsilon > 0$ is small but nonzero (*i.e.* in the real situation where fecundity is large but finite)? Numerical results in Section 10.3.2 begin to answer this challenging question.

**Remark 10.2.** *It is notable that the proof of Theorem 10.1 rests upon a property of the functions involved in Equation (10.15), namely the existence of a unique $p^* \in (0,1)$ such that $(f/g)'(p^*) = 0$, and $C \neq -Tf(p^*)/g(p^*)$. Indeed, letting $\xi = \frac{d_1 b_2^0}{d_2 b_1^0}$ we have*

$$\frac{f}{g}(p) = Kd_2\big(\frac{p}{1-s_h p}\xi - p\big), \quad \big(\frac{f}{g}\big)'(p) = Kd_2\big(\frac{1}{(1-s_h p)^2}\xi - 1\big).$$

*The roots of the second-order polynomial at the numerator of the right-hand side read*

$$p_\pm = \frac{1}{s_h}\big(1 \pm \sqrt{\xi}\big),$$

*so assuming (10.6) (i.e. $\xi < 1$) yields*

$$p^* = \frac{1}{s_h}\big(1 - \sqrt{\xi}\big)$$

*(which indeed belongs to $[0,\theta)$ as a consequence of (10.6): from $1 - s_h < \xi < 1$ it follows that $0 < p^* < (1-\sqrt{1-s_h})/s_h < 1$ since $s_h \in (0,1]$). On the contrary, assuming $d_2 b_1^0 < d_1 b_2^0$ (i.e. $\xi > 1$) implies that there is no such $p^*$ in $[0,1]$ (and in this case the control must be bang-bang, as a consequence of Lemma 10.8 below).*

From Proposition 10.2 and Theorem 10.1, we provide hereafter a more precise result about the convergence of optimal values for Problem ($\mathcal{P}_{\text{full}}$) as $\epsilon \searrow 0$.

**Corollary 10.1.** *Let $(u^\epsilon)_{\epsilon>0}$ be a family of minimizers for Problem ($\mathcal{P}_{\text{full}}$). Then, $(u^\epsilon)_{\epsilon>0}$ converges strongly in $L^1(0,T)$ to a solution of Problem ($\mathcal{P}_{\text{reduced}}$) as $\epsilon \searrow 0$ (which is unique whenever $C \neq C^*(M)$ with the notations of Theorem 10.1).*

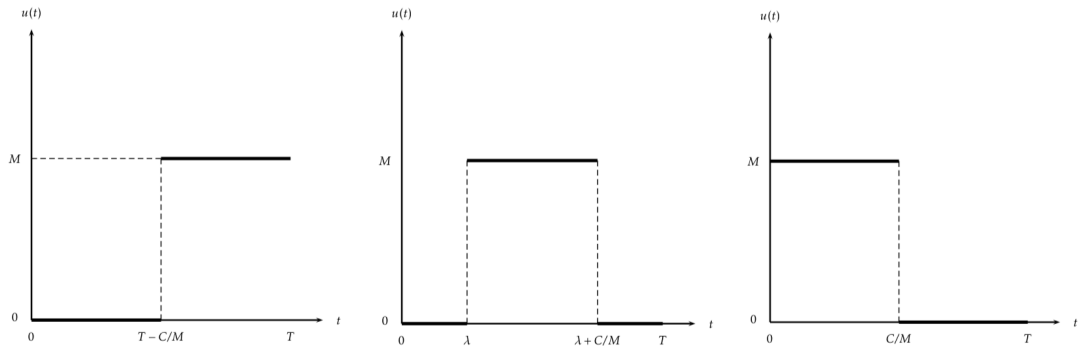



Figure 10.3: Left: solution $u^*$ in the case $M > \max_{p\in[0,\theta]} -f(p)/g(p)$ and $C > C^*(M)$. Middle: one solution $u^*_\lambda$ in the case $M > \max_{p\in[0,\theta]} -f(p)/g(p)$ and $C = C^*(M)$. Right: solution $u^*$ in the case $M \leq \max_{p\in[0,\theta]} -f(p)/g(p)$ or $M > \max_{p\in[0,\theta]} -f(p)/g(p)$ and $C < C^*(M)$. .

### 10.3.2 Numerics

This section is devoted to computing the solution of Problem ($\mathcal{P}_{\text{full}}$) and to illustrating the relations with its reduced version ($\mathcal{P}_{\text{reduced}}$).

All the simulations are obtained with a direct method applied to the optimal control problem ($\mathcal{P}_{\text{full}}$), consisting in discretizing System (10.1), the control, and to reduce the optimal control problem to some minimization problem with constraints. To this aim, we used the open-source optimization routine from `IPOPT` (see [231]) combined with `AMPL` modeling language (see [89]). This enables the computation of a local minimizer for a discretized version of ($\mathcal{P}_{\text{full}}$).

**Choice of numerical parameters and methods.** Populations are normalized by setting $K = 1$, and Table 10.1 yields the values used for the other parameters. The time-dynamics (the slow-fast system (10.12) depending on $\epsilon$) are discretized with the Runge-Kutta implicit scheme Lobatto IIIC of order 2 (two stages). This scheme is asymptotic preserving in $\epsilon$ (see [98]) and allows for sound comparison of the simulations across a range of values of this parameter.

We obtain a solution $\mathbf{n}_{\Delta t} \in (\mathbb{R}_+)^{2N_d}$ as well as an approximate local minimizer for the discretized problem ($\mathcal{P}_{\text{full}}$), $\widehat{u}^{\epsilon,\Delta t} \in [0, M]^{N_d}$.

| *Category* | *Parameter* | *Name* | *Value or range* |
|---|---|---|---|
| Discretization | $\Delta t$ | Time step | $[0.0004, 0.0015]$ |
| Singular limit | $1/\epsilon$ | Birth rates normalization | $[1, 2000]$ |
| Optimization | $T$ | Final time | 10 |
| | $C$ | Maximal release number | $[0.15, 0.75]$ |
| | $M$ | Maximal release flux | 10 |
| Biology | $b_1^0$ | Normalized wild birth rate | 1 |
| | $b_2^0$ | Normalized infected birth rate | 0.9 |
| | $d_1$ | Wild death rate | 0.27 |
| | $d_2$ | Infected death rate | 0.3 |
| | $s_h$ | Cytoplasmic incompatibility level | 0.9 |

Table 10.1: Parameters for the numerical resolution of ($\mathcal{P}_{\text{full}}$)

**Results.** It is convenient to introduce the number of steps in the time discretization $N_d = T/\Delta t$.

For the parameters given in Table 10.1, we can compute the critical value $C^*(M)$ from Theorem 10.1 numerically: it is close to 0.24. Therefore we choose three values of $C$ (0.15, 0.4 and 0.75) both above and below this threshold, so as to get contrasting results. On Figure 10.4 below, solutions of Problem ($\mathcal{P}_{\text{full}}$) are computed for these three different values of the integral bound $C$ and for $\epsilon = 1$. We observe that the set $I_M^{\Delta t,\kappa} := \{k \in [\![1, N_d]\!],\ \widehat{u}_k^{\epsilon,\Delta t} \geq M - \kappa\}$ (approximating the set $\{u = M\}$), for $\kappa$ small enough, is made of two segments containing either 1 or $N_d$. Let us denote these two segments $[\![1, k_0(\Delta t)]\!]$ and $[\![k_1(\Delta t), N_d]\!]$. It seems that

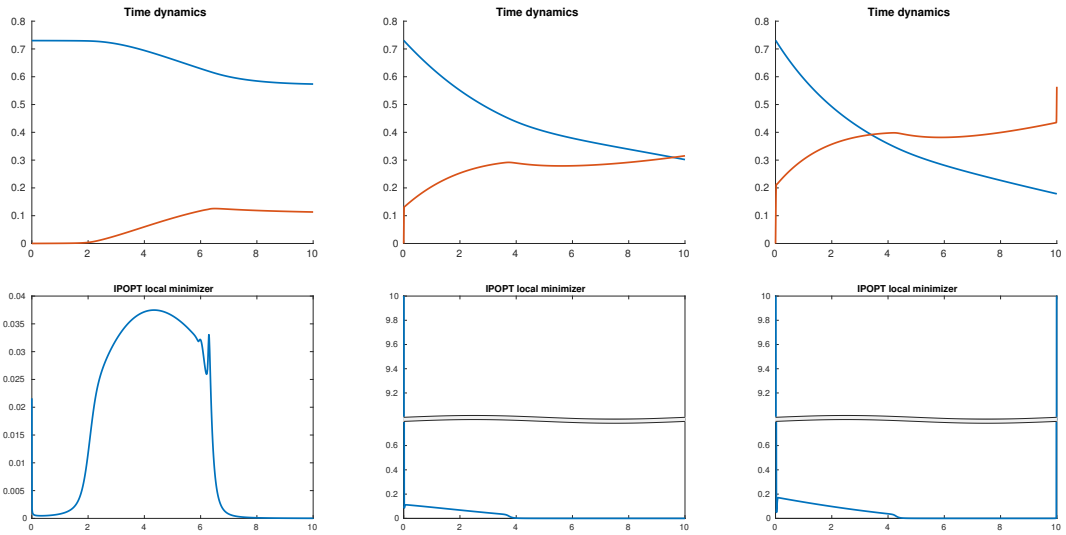



Figure 10.4: Top: time dynamics (plots of the wild mosquitoes density $n_1$ starting from a positive value *versus* the *Wolbachia*-infected mosquitoes density $n_2$ starting from 0). Bottom: numerical optimal control. From the left to the right: $C = 0.15$, $C = 0.4$ and $C = 0.75$. The parameter $\epsilon$ is fixed to 1.

- a *relaxation type* phenomenon may occur for optimal controls meaning that the solution is not *bang-bang*.
- the set $I_{\text{relax}}^{\Delta t,\kappa} := \{k \in [\![1, N_d]\!],\ \kappa \leq \widehat{u}_k^{\epsilon,\Delta t} \leq M - \kappa\}$ (approximating the set $\{0 < u < M\}$) seems to be a segment for $\kappa$ small enough.
- for small values of $C$, $k_0 = 0$, $k_1 = N_d$ and there is replacement failure, suggesting that it is necessary to release a minimal number of infected mosquitoes in order to guarantee population replacement.

Figures 10.5 and 10.6 are used to validate our approach of considering the asymptotic problem ($\mathcal{P}_{\text{reduced}}$) instead of the real one ($\mathcal{P}_{\text{full}}$), with $C = 0.75$ (leading to replacement success) and $C = 0.15$ (leading to replacement failure), respectively. We compare the numerical values of $J(u = \widehat{u}^{\epsilon,\Delta t})$ obtained either by using the direct optimization routine described above, or by choosing $u = u_0^*$ as the (explicit) solution of Problem ($\mathcal{P}_{\text{reduced}}$). As expected, the ratio

$$\frac{J^\epsilon(u_0^*) - J^\epsilon(\widehat{u}^{\epsilon,\Delta t})}{J^\epsilon(\widehat{u}^{\epsilon,\Delta t})}$$

visually converges to 0 as $\epsilon \searrow 0$. The bottom panels in figures 10.5 and 10.6 illustrate the convergence properties for $p^\epsilon$ stated in Proposition 10.1, and for $u^\epsilon$ stated in Corollary 10.1.

## 10.4 Conclusion

In this article, we proposed a strategy of *Wolbachia*-infected mosquitoes releases to control a simplified competitive compartmental system involving wild and infected individuals. Our approach is validated by numerical results that seem promising. Hereafter, we enumerate a list of issues that remain open and will be investigated in a future work.

**Partial or complete solving of Problem ($\mathcal{P}_{\text{full}}$).** When investigating numerically this problem (see Section 10.3.2), we observed several interesting properties of minimizers, at least for several relevant values of parameters: *relaxation* phenomena may appear (meaning that the minimizer $u^*$ is not *bang-bang* anymore). The set $\{u^* = M\}$ seems to have two connected components meeting 0 and $T$. Some qualitative properties of the solutions to ($\mathcal{P}_{\text{full}}$) are discussed in Appendix 10.B.

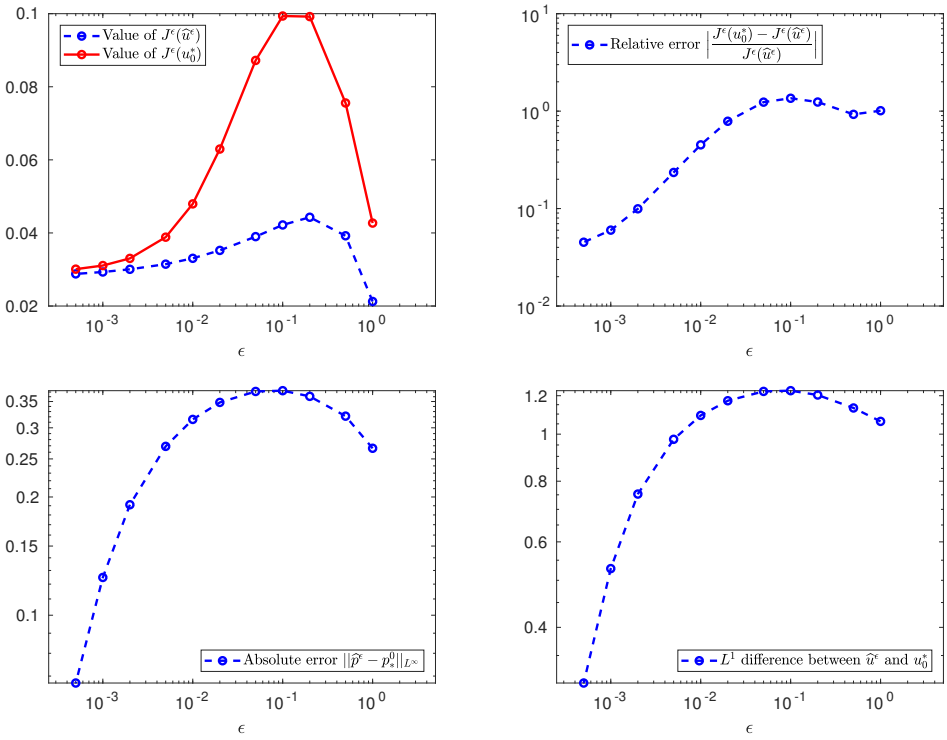



Figure 10.5: Case $C = 0.75$. Top left: numerical minimum value $J^\epsilon(\widehat{u}^{\epsilon,\Delta t})$ and $J^\epsilon(u_0^*)$ w.r.t. $\epsilon$. Top right: relative error between the value of $J^\epsilon$ at the numerical minimizer $\widehat{u}^{\epsilon,\Delta t}$ and at the exact solution $u^*$ of the asymptotic problem ($\mathcal{P}_{\text{reduced}}$) w.r.t. $\epsilon$. Bottom left: plot of the absolute error between $\widehat{p}^\epsilon$ and $p_0^*$. Bottom right: $L^1$ error between $\widehat{u}^{\epsilon,\Delta t}$ and $u_0^*$

**Asymptotic of Problem ($\mathcal{P}_{\text{full}}$) when one makes simultaneously $\epsilon$ go to zero and $M$ (the pointwise upper-bound constraint on $u$) go to $+\infty$.** According to Theorem 10.1, one shows easily that making successively $\epsilon$ tend to 0 and then $M$ tend to $+\infty$ yields to a new asymptotic problem whose minimizers are a (typically unique) Dirac mass. When making simultaneously $\epsilon$ tend to 0 and then $M$ tend to $+\infty$, the behavior of minimizers is not so clear and a careful analysis must be led to understand it. We refer to Section 13.3 for further discussion on this topic.

**Investigation of a more realistic model.** Coming back to the initial *Wolbachia*-infected mosquitoes control problem, it is likely that a model taking into account dispersal effects would provide more satisfying and workable results. To this aim, System (10.1) could be replaced by a more general reaction-diffusion system of partial differential equations. It is likely that numerical difficulties may arise for the related optimization problem, needing to develop an adapted approach.

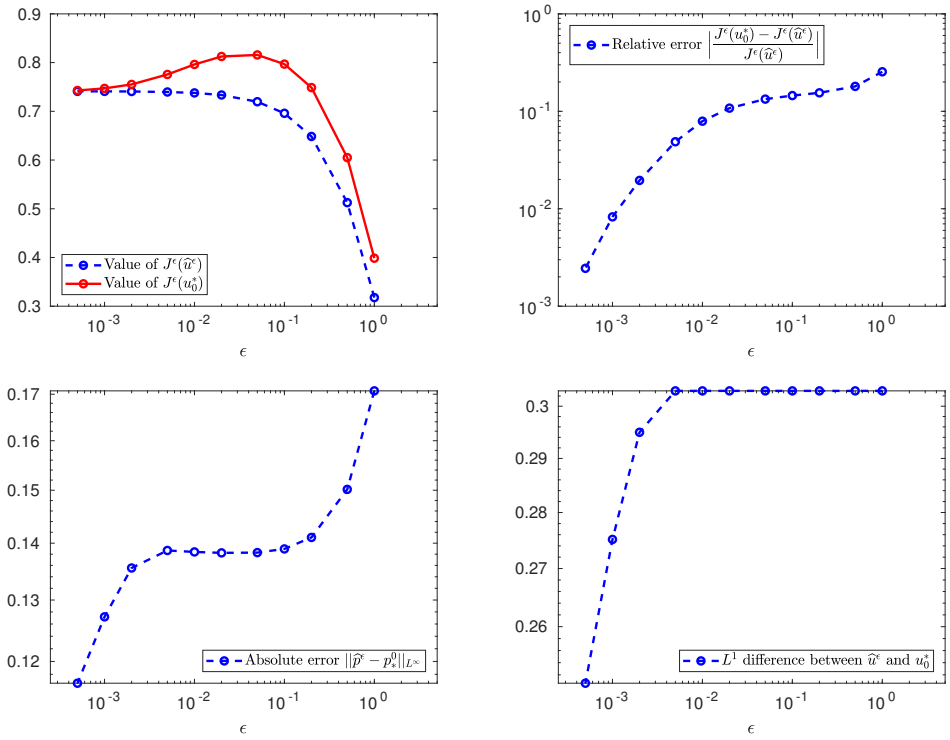



Figure 10.6: Case $C = 0.15$. Top left: numerical minimum value $J^\epsilon(\widehat{u}^{\epsilon,\Delta t})$ and $J^\epsilon(u_0^*)$ w.r.t. $\epsilon$. Top right: relative error between the value of $J^\epsilon$ at the numerical minimizer $\widehat{u}^{\epsilon,\Delta t}$ and at the exact solution $u^*$ of the asymptotic problem ($\mathcal{P}_{\text{reduced}}$) w.r.t. $\epsilon$. Bottom left: plot of the absolute error between $\widehat{p}^\epsilon$ and $p_0^*$. Bottom right: $L^1$ error between $\widehat{u}^{\epsilon,\Delta t}$ and $u_0^*$

# Appendices

## 10.A Proofs

### 10.A.1 Proof of Lemma 10.1

Solving the equation $\mathbf{f}(n_1, n_2) = 0$ yields the steady states by direct computation. Let us use the notations $N = (n_1 + n_2)/K$ and $p = n_2/(n_1 + n_2)$. The Jacobian associated to the right-hand side $\mathbf{f}$ of the system reads

$$\mathbf{Jac}(\mathbf{n}) = \begin{pmatrix} b_1\big((1-s_h p)(1-(2-p)N) + s_h p(1-p)(1-N)\big) - d_1 & -b_1(1-p)\big(s_h(1-p) + N(1-s_h)\big) \\ -b_2 pN & b_2(1-(1+p)N) - d_2 \end{pmatrix}.$$

It is readily seen that the extra-diagonal terms are non-positive (and even negative if $p \in (0,1)$ and $N > 0$). By Kamke-Muller conditions (see [114]), this implies that the system is monotone with respect to the cone $\mathbb{R}_+ \times \mathbb{R}_-$, in other words it is competitive.

In particular,

$$\mathbf{Jac}(n_1^*, 0) = \begin{pmatrix} -(b_1 - d_1) & -b_1 + (1-s_h)d_1 \\ 0 & b_2 d_1/b_1 - d_2 \end{pmatrix},$$
$$\mathbf{Jac}(0, n_2^*) = \begin{pmatrix} b_1 d_2(1-s_h)/b_2 - d_1 & 0 \\ -(b_2 - d_2) & -(b_2 - d_2) \end{pmatrix},$$

so that conditions (10.5) and (10.6) easily yield the linear stability of $(n_1^*, 0)$ and $(0, n_2^*)$. Combined with the monotonicity property of the system, we get the asymptotic stability.

Then, $\mathbf{n}^C$ belongs to the interior of the interval $[(n_1^*, 0), (0, n_2^*)]$ (for the order induced by the comparison principle recalled in Footnote 2), whose bounds are stable steady states, and there is no other steady state in the interior of this interval. Hence it must be unstable, since the dynamics of (10.1) is order-preserving.

At $(0,0)$, we compute the directional derivative in direction $(h,k)$ as

$$D\mathbf{f}(h,k) = \lim_{t \to 0} \frac{\mathbf{f}(th, tk)}{t} = \begin{pmatrix} (b_1(1 - s_h \frac{k}{h+k}) - d_1)h \\ (b_2 - d_2)k \end{pmatrix},$$

and in particular we find that the direction $(0,1)$ is unstable.

### 10.A.2 Proof of Proposition 10.1

System (10.1) reads

$$\begin{cases} \dfrac{dn_1^\epsilon}{dt} = & b_1^0 n_1^\epsilon (1 - s_h p^\epsilon) n^\epsilon - d_1 n_1^\epsilon \\ \dfrac{dn_2^\epsilon}{dt} = & b_2^0 n_2^\epsilon n^\epsilon - d_2 n_2^\epsilon + u^\epsilon. \end{cases} \tag{10.20}$$

Hence, the resulting system (10.12) on $(n^\epsilon, p^\epsilon)$ in Proposition 10.1 is obtained from straightforward computations.

Let us now provide *a priori* bounds on $(n^\epsilon(t), p^\epsilon(t))$ (uniform in $\epsilon \leq \epsilon_0$, for all $t \geq 0$). Note that $0 \leq p^\epsilon \leq 1$ is an easy consequence of the Cauchy-Lipschitz theorem since $p^\epsilon = 0$ and $p^\epsilon = 1$ are respectively sub- and super-solutions.

We infer that the right-hand side of the equation on $n^\epsilon$ in (10.12) is bounded from below by

$$a(p)(1-\epsilon n)(Z(p)-n)-\frac{M}{K},$$

which is positive as soon as $n$ is smaller than the smallest root of this second order polynomial in $n$ given by

$$\frac{a(p)(1+\epsilon Z(p))-a(p)\sqrt{(1-\epsilon Z(p))^2+4\epsilon M/(Ka(p))}}{2a(p)\epsilon}.$$

Moreover, the right-hand side of the equation on $n^\epsilon$ in (10.12) is bounded from above by

$$a(p)(1-\epsilon n)(Z(p)-n),$$

which is negative as soon as $n$ is between $Z(p)$ and $1/\epsilon$. We then infer the expected uniform estimates on $n^\epsilon$ as soon as $\epsilon_0$ is small enough.

We are then driven to the slow-fast system (10.12). Using the uniform bounds on $n^\epsilon$, $p^\epsilon$, $u^\epsilon$, we infer that the right-hand sides are bounded. Hence, by using the Arzelà-Ascoli theorem, we get that $(p^\epsilon)_{\epsilon>0}$ converges up to a subfamily uniformly to some function $p$ such that $p(0)=0$ and $0\le p\le 1$ as $\epsilon\searrow 0$. Moreover, $dp/dt$ is uniformly bounded since $dp^\epsilon/dt$ is.

**Lemma 10.2.** *Up to a subfamily, the family $(n^\epsilon)_{\epsilon>0}$ converges weakly to $Z(p)-\frac{u}{K}$ as $\epsilon\searrow 0$ in $(W^{1,1})'$, with $p$ the uniform limit of any subfamily $(p^\epsilon)_{\epsilon>0}$.*

*Proof.* Let $\phi\in W^{1,1}$ and multiply the differential equation satisfied by $n^\epsilon$ by $\phi$ and integrate by parts over $[0,T]$. We get

$$\epsilon[\phi n^\epsilon]_0^T-\epsilon\int_0^T\frac{d\phi}{dt}n^\epsilon=\int_0^T\phi a(p^\epsilon)(Z(p^\epsilon)-n^\epsilon)-\int_0^T\phi\frac{u^\epsilon}{K}.$$

By weak-star convergence of $u^\epsilon$ in $L^\infty$, uniform convergence of $p^\epsilon$ in $L^\infty$ and uniform boundedness of $n^\epsilon$ we infer that

$$0=\lim_{\epsilon\to 0}\int_0^T\phi a(p^\epsilon)(Z(p^\epsilon)-n^\epsilon)-\int_0^T\phi\frac{u}{K},$$

leading to the expected result. □

**Lemma 10.3.** *Up to a subfamily, $(p^\epsilon)_{\epsilon>0}$ converges uniformly to $p$ solving the ordinary differential equation*

$$\frac{dp}{dt}=\beta(n,p,u),\quad p(0)=0.$$

*Proof.* Let us first recast the equation on $p^\epsilon$ in system (10.12) under the form

$$\frac{dp^\epsilon}{dt}=\beta_\epsilon(n^\epsilon,p^\epsilon,u^\epsilon),\quad p^\epsilon(0)=0,$$

with

$$\beta_\epsilon(n,p,u)=p(1-p)\big(n(b_2^0-b_1^0(1-s_hp))+d_1-d_2\big)+\frac{u(1-p)}{K(1-\epsilon n)}$$

so that we easily infer (with obvious notations) that $\beta_\epsilon\to\beta$ as $\epsilon\searrow 0$ with $\beta_\epsilon(n,p,u)=n\widehat{\beta}(p)+\beta_0(p)+u\widetilde{\beta}^\epsilon(n,p)$ and $\beta(n,p,u)=n\widehat{\beta}(p)+\beta_0(p)+u\widetilde{\beta}(p)$.

Using the previous considerations (and in particular the uniform boundedness of $p^\epsilon$, $n^\epsilon$ and $u^\epsilon$), we deduce that $p$ is in fact Lipschitz-continuous, and that $\widehat{\beta}$ and $\widetilde{\beta}$ are continuous on $[0,1]$.

Now, let us show that $p$ satisfies the limit equation in a weak sense. Let $\phi\in\mathcal{C}_c^\infty(0,T)$. We compute each term separately: the terms in $dp^\epsilon/dt$ and $\beta_0(p^\epsilon)$ converge by uniform convergence of $p^\epsilon$. Therefore, we have

$$\int_0^T\phi n^\epsilon\widehat{\beta}(p^\epsilon)=\underbrace{\int_0^T\phi n^\epsilon\widehat{\beta}(p)}_{\to\int_0^T\phi n\widehat{\beta}(p)}+\underbrace{\int_0^T\phi n^\epsilon\big(\widehat{\beta}(p^\epsilon)-\widehat{\beta}(p)\big)}_{|\cdot|\le\|n^\epsilon\|_\infty o(1)}.$$

and

$$\int_0^T \phi u^\epsilon \widetilde{\beta}^\epsilon(n^\epsilon, p^\epsilon) = \underbrace{\int_0^T \phi u^\epsilon \widetilde{\beta}(p)}_{\to \int_0^T \phi u \widetilde{\beta}(p)} + \underbrace{\int_0^T \phi u^\epsilon \left(\widetilde{\beta}^\epsilon(n^\epsilon, p^\epsilon) - \widetilde{\beta}(p)\right)}_{|\cdot| \leq M o(1)}.$$

by using simultaneously the weak convergence properties of $(u^\epsilon)_{\epsilon>0}$ and $(n^\epsilon)_{\epsilon>0}$ (see Lemma 10.2) as well as the aforementioned convergence of $\beta_\epsilon$ to $\beta$. Here, it is crucial that the limit $\widetilde{\beta}$ does not depend on $n$ but merely on $p$, and we rely on the uniform estimate on $n^\epsilon$.

Finally a standard argument yields that $p$ must satisfy the equation in a strong sense since it is Lipschitz-continuous. □

We are now in position to conclude the proof of Proposition 10.1. Indeed, the limit $p$ satisfies (with obvious notations) an equation of the form

$$\frac{dp}{dt} = f(p) + ug(p).$$

Since the solution to this equation is unique, we finally get the uniform convergence of the whole family $(p^\epsilon)_{\epsilon>0}$ to $p$.

### 10.A.3 Proof of Proposition 10.2

Let us first investigate the existence of solutions for Problem ($\mathcal{P}_{\text{full}}$) under the assumption (10.9).

Fix $\epsilon > 0$ and consider $(u_n^\epsilon)_{n\in\mathbb{Z}_{\geq 0}}$ a minimizing sequence. According to the Banach-Alaoglu Bourbaki theorem, the set $\mathcal{U}_{T,C,M}$ is compact for the weak star topology of $L^\infty(0,T)$. Therefore, up to a subsequence, $(u_n^\epsilon)_{n\in\mathbb{Z}_{\geq 0}}$ converges to some element $u^\epsilon \in \mathcal{U}_{T,C,M}$. Let us use the same notation to denote $(u_n^\epsilon)_{n\in\mathbb{Z}_{\geq 0}}$ and any converging subsequence (with a slight abuse of notation).

An immediate adaptation of the proof of Proposition 10.1 yields successively that $(\mathbf{n}_n^\epsilon)_{n\in\mathbb{Z}_{\geq 0}}$ (the sequence of solutions $\mathbf{n}_n^\epsilon$ of System (10.1) corresponding to $u = u_n^\epsilon$) is uniformly bounded and converges uniformly to some limit $\mathbf{n}^\epsilon$ as $n \to +\infty$, which corresponds to the solution of system (10.1) with $u = u^\epsilon$. We then infer that $(J^\epsilon(u_n^\epsilon))_{n\in\mathbb{Z}_{\geq 0}}$ converges to $J^\epsilon(u^\epsilon)$ and the conclusion follows.

To prove the convergence of minimizers as $\epsilon \searrow 0$ and the existence of solutions for Problem ($\mathcal{P}_{\text{reduced}}$), we will show that $J^\epsilon$ $\Gamma$-converges[4] to $J^0$ as $\epsilon \to 0$, and conclude by using for instance [39, Theorem 2.1].

We compute

$$J^\epsilon(u^\epsilon) = \frac{K^2}{2}\Big((1-\epsilon n^\epsilon(T))^2(1-p^\epsilon(T))^2 + \big(1-p^\epsilon(T) - \epsilon(\frac{d_2}{b_2^0} - p^\epsilon(T)n^\epsilon(T))\big)_+^2\Big)$$

so that if $\epsilon > 0$ is small enough,

$$J^\epsilon(u^\epsilon) = \frac{K^2}{2}\Big(2(1-p^\epsilon(T))^2 - 2\epsilon n^\epsilon(T)(1-p^\epsilon(T))(2p^\epsilon(T)-1) - 2\epsilon(1-p^\epsilon(T))\frac{d_2}{b_2^0} + \epsilon^2(\frac{d_2}{b_2^0})^2 - 2\epsilon^2 p^\epsilon(T)n^\epsilon(T) + \epsilon^2 n^\epsilon(T)^2((1-p^\epsilon(T))^2 + p^\epsilon(T)^2)\Big).$$

By the uniform estimates on $n^\epsilon, p^\epsilon$ provided in Proposition 10.1, we get that $dp^\epsilon/dt$ is uniformly bounded in $\epsilon$ on $[0,T]$, and thus by Arzelà-Ascoli theorem up to extraction $p^\epsilon$ converges uniformly to some $p$.

In the particular case where $\overline{u}^\epsilon = u$ we get that $\lim_{\epsilon\to 0} J^\epsilon(u) = J^0(u)$, which implies (10.22). Indeed, according to Proposition 10.1, the limit $p$ is unique and solves precisely

$$\frac{dp}{dt} = f(p) + ug(p), \quad p(0) = 0.$$

[4] Reminder about $\Gamma$-convergence: one says that $J^\epsilon$ $\Gamma$-converges to $J^0$ if for $u \in \mathcal{U}_{T,C,M}$ and $(u^\epsilon)_{\epsilon>0}$ converging weak-star to $u$ in $L^\infty(0,T)$, one has

$$\liminf_{\epsilon\to 0} J^\epsilon(u^\epsilon) \geq J^0(u) \tag{10.21}$$

and there exists a sequence $(\overline{u}^\epsilon)_\epsilon$, with $\overline{u}^\epsilon \rightharpoonup u$, such that

$$\limsup_{\epsilon\to 0} J^\epsilon(\overline{u}^\epsilon) \leq J^0(u). \tag{10.22}$$

Note that Proposition 10.1 proves in fact the stronger result that $J^\epsilon(u^\epsilon)$ converges to $J^0(u)$ as $\epsilon$ goes to 0, whence (10.21).

### 10.A.4 Proof of Theorem 10.1

The proof relies on several intermediary lemmas which we state and prove below. We first prove that the $L^1$ constraint on the control $u$ is saturated.

**Lemma 10.4.** *If $u^*$ solves the optimization problem ($\mathcal{P}_{\text{reduced}}$), then* $\displaystyle\int_0^T u^*(t)\,dt = \min(C, TM)$.

*Proof.* This is a consequence of the fact that the function $g$ defined by (10.17) satisfies $g(p) > 0$ for $p \in [0, 1)$. Indeed, if $u_1 < u_2$ then if there exists a time $\tau \geq 0$ such that the corresponding solution to (10.4) satisfies $q_1(\tau) = q_2(\tau) < 1$. We deduce from (10.4) that $\dot{q_1}(\tau) < \dot{q_2}(\tau)$. Thus there exists $\tau_1 > \tau$ such that $q_1(t) < q_2(t)$ on $[\tau, \tau_1]$.

As a consequence, if $h$ is a nonnegative function such that $\int_{[0,T]} h > 0$ then $J^0(u+\alpha h) < J^0(u)$ whenever $\alpha$ is small enough. Thus, the constraint is saturated. □

Let us define the adjoint state $q$ defined by

$$-\dot{q} = \big(f'(p) + ug'(p)\big)q \text{ on } (0,T), \qquad q(T) = -2(1-p(T)). \tag{10.23}$$

Standard arguments yield existence and uniqueness of a solution for System (10.23). Moreover, since $0 < p(\,\cdot\,) < 1$, we deduce that $q(\,\cdot\,) < 0$ on $[0, T]$.

Let us now state the (necessary) first order optimality conditions for Problem ($\mathcal{P}_{\text{reduced}}$).

**Lemma 10.5.** *Let $u \in \mathcal{U}_{T,C,M}$. Then, for every admissible perturbation[5] $h$, the Gâteaux-derivative of $J^0$ at $u$ in the direction $h$ reads*

$$\langle dJ^0(u), h\rangle = \int_0^T h(t)q(t)g(p(t))dt.$$

*Proof.* Let $h$ be an admissible perturbation of $u$ (see Footnote 5). The Gâteaux-differentiability of $J^0$ is standard and follows from the differentiability of the mapping $\mathcal{U}_{T,C,M} \ni u \mapsto p$, where $p$ denotes the unique solution of (10.15), itself deriving from the application of the implicit functions theorem combined with variational arguments.

Let us then compute the Gâteaux-derivative of $J^0$ at $u$ in the direction $h$, defined by

$$\langle dJ^0(u), h\rangle = \lim_{\varepsilon\to 0} \frac{J^0(u+\varepsilon h) - J^0(u)}{\varepsilon}.$$

Let us introduce $\delta p$, the Gâteaux-differential of $p$ at $u$ in the direction $h$. Straightforward computations yield that $\delta p$ solves the linearized problem to (10.4),

$$\dot{\delta p}(t) = f'(p)\delta p + ug'(p)\delta p + hg(p), \qquad \delta p(0) = 0.$$

Then, one has

$$\langle dJ^0(u), h\rangle = -2(1-p(T))\delta p(T) = q(T)\delta p(T),$$

where $q$ is the solution to the adjoint equation (10.23). Then, we compute

$$0 = \int_0^T \delta p(\dot{q} + f(p)q + ug'(p)q)\,dt = \delta p(T)q(T) - \delta p(0)q(0) - \int_0^T h(t)q(t)g(p(t))\,dt$$

and we infer that

$$\langle dJ^0(u), h\rangle = \int_0^T h(t)q(t)g(p(t))\,dt.$$

□

[5] More precisely, we call "admissible perturbation" any element of the tangent cone $\mathcal{T}_{u,\mathcal{U}_{T,C,M}}$ to the set $\mathcal{U}_{T,C,M}$ at $u$. The cone $\mathcal{T}_{u,\mathcal{U}_{T,C,M}}$ is the set of functions $h \in L^\infty(0,T)$ such that, for any sequence of positive real numbers $\varepsilon_n$ decreasing to 0, there exists a sequence of functions $h_n \in L^\infty(0,T)$ converging to $h$ as $n \to +\infty$, and $u + \varepsilon_n h_n \in \mathcal{U}_{T,C,M}$ for every $n \in \mathbb{Z}_{\geq 0}$ (see e.g. [109, chapter 7]).

**Lemma 10.6.** *Let $u \in \mathcal{U}_{T,C,M}$ be a solution of Problem ($\mathcal{P}_{\text{reduced}}$). Define the switching function $w$ by $w(t) = g(p(t))q(t)$ for all $t \in [0,T]$. There exists $\Lambda < 0$ such that*

- $u(t) = M \iff w(t) < \Lambda$,
- $0 < u(t) < M \iff w(t) = \Lambda$,
- $u(t) = 0 \iff w(t) > \Lambda$,

*each equality being understood up to a zero Lebesgue-measure set.*

*Proof.* Introduce the Lagrangian function $\mathcal{L}$ associated to Problem ($\mathcal{P}_{\text{reduced}}$), defined by

$$\mathcal{L} : \mathcal{U}_{T,C,M} \times \mathbb{R} \ni (u, \Lambda) \mapsto J^0(u) - \Lambda \left( \int_0^T u(t)\, dt - C \right).$$

Standard arguments enable to show the existence of a Lagrange multiplier $\Lambda$ such that $(u, \Lambda)$ is a saddle-point of the Lagrangian functional $\mathcal{L}$. Moreover, according to Lemma 10.4 and since $T > C/M$, we have necessarily $\int_0^T u = C$.

Let $x_0$ be a density-one point of $\{u = M\}$. Let $(G_{k,n})_{n \in \mathbb{Z}_{\geq 0}}$ be a sequence of measurable subsets with $G_{n,k}$ included in $\{u = M\}$ and containing $x_0$. Let us consider $h = \mathbb{1}_{G_{k,n}}$ and notice that $u - \eta h$ belongs to $\mathcal{U}_{T,C,M}$ whenever $\eta$ is small enough. Writing

$$\mathcal{L}(u - \eta h, \Lambda) \geq \mathcal{L}(u, \Lambda),$$

dividing this inequality by $\eta$ and letting $\eta$ go to 0, it follows that

$$-\langle dJ^0(u), h \rangle + \Lambda \int_0^T h(t)\, dt \geq 0$$

or equivalently that

$$-\int_{G_{n,k}} q(t)g(p(t)) + \Lambda |G_{n,k}| \geq 0.$$

according to Lemma 10.5. Dividing this inequality by $|G_{k,n}|$ and letting $G_{k,n}$ shrink to $\{x_0\}$ as $n \to +\infty$ shows the first point of Lemma 10.6, according to the Lebesgue Density Theorem. The proof of the third point is similar, and consists in considering perturbations of the form $u + \eta h$ where $h$ denotes a positive admissible perturbation of $u$ supported in $\{u(t) = 0\}$. Finally, the proof of the second point follows the same lines, by considering bilateral perturbations of the form $u \pm \eta h$ where $h$ denotes an admissible perturbation of $u$ supported in $\{0 < u(t) < M\}$.

Note also that the obtained properties are in fact equivalent by observing that the sets $\{w(t) < \Lambda\}$, $\{w(t) > \Lambda\}$ and $\{w(t) = \Lambda\}$ realize a partition of $[0,T]$. □

**Lemma 10.7.** *Let $u \in \mathcal{U}_{T,C,M}$ be a solution of Problem ($\mathcal{P}_{\text{reduced}}$). One has $0 < u(t) < M$ on an open interval containing $t$ if and only if $w'(t) = 0$, which rewrites $f'(p(t))g(p(t)) = f(p(t))g'(p(t))$.*

*Under the assumption (10.6), there exists a unique $p^* \in (0,1)$ such that $(f/g)'(p^*) = 0$ and therefore, $\{u \in (0,M)\}$ is an open interval containing $t$ if and only if $p(t) = p^*$, which implies $u(t) = -f(p^*)/g(p^*)$.*

*Proof.* Let us differentiate $t \mapsto g(p(t))q(t)$. We get

$$\begin{aligned} \frac{d}{dt}\big(q(t)g(p(t))\big) &= q'g + p'g'q \\ &= (-f' - ug')gq + (f + ug)g'q \\ &= q(t)\big(f(p(t))g'(p(t)) - f'(p(t))g(p(t))\big). \end{aligned}$$

Combining this computation with Remark 10.2 yields the expected result. □

From this general fact we deduce

**Lemma 10.8.** *Let $u \in \mathcal{U}_{T,C,M}$ be a solution of Problem ($\mathcal{P}_{\text{reduced}}$). Under the assumption (10.6) and if $M > \max_{[0,1]} -f/g$, $u$ is either bang-bang or constant and equal to $-f(p^*)/g(p^*)$ (the latter case may occur only if $C = -Tf(p^*)/g(p^*)$).*

*Proof.* Between 0 and 1, $f$ changes sign only once, at $\theta$. In addition, the switching function $w : t \mapsto q(t)g(p(t))$ is decreasing if $p(t) < p^*$ and increasing if $p(t) > p^*$, since it is positively proportional to $(f/g)'$, which changes sign only once, and $f/g$ changes sign only once, and is decreasing at 0, so $(f/g)'$ has the same sign as $p - p^*$.

Necessarily, $\theta \geq p^*$. Indeed, $f/g$ is decreasing on $(0, p^*)$ and equal to 0 at 0 and $\theta$.

Let $I = (t_1, t_2)$ be the maximal interval on which $p(t) = p^*$, $u(t) = -f(p^*)/g(p^*)$, $w(t) = \Lambda$. If at $t_2^+$ we have $u = 0$ then $p$ must decrease since $p^* < \theta$, so $p(t) < p^*$ at $t_2^+$, and therefore $w$ must decrease at $t_2$, but this contradicts the necessary optimality condition of Lemma 10.6. If at $t_2^+$ we have $u = M$ then $p$ must increase if $M$ is large enough. Then $w$ must increase, and again this is in contradiction with Lemma 10.6. Hence $I = \emptyset$ or $I = [0, T]$. But $I = [0, T]$ is admissible if and only if $-Tf(p^*)/g(p^*) = C$. □

Let us define $p_M$ as the solution of

$$\frac{dp_M}{dt} = f(p_M) + Mg(p_M), \quad p_M(0) = 0.$$

Assume that $M > \max_{p\in[0,\theta]} -\frac{f(p)}{g(p)}$. Then $\frac{dp_M}{dt} = f(p_M) + Mg(p_M) > 0$. Introduce the function $G_M$ defined by $G_M'(p) = \frac{1}{f(p)+Mg(p)}$ and $G_M(0) = 0$. Then, $G_M$ is an increasing function and we have

$$G_M(p_M(t)) = G_M(p_M(t_0)) + t - t_0, \quad \text{and} \quad G_M(p_M(C/M)) = \frac{C}{M}.$$

The use of all these results allows us to conclude the proof of Theorem 10.1.

*Proof of Theorem 10.1.* We split the proof into three cases :

- *Case* $p_M(C/M) < \theta$. This condition is equivalent to $G_M(p_M(C/M)) < G_M(\theta)$ (since $G_M$ is increasing). By Lemma 10.8, the control $u$ is bang-bang and the set where $u = M$ is open, (since from Lemma 10.6, it is the set of interval on which $g(p)q < \Lambda$). Consider that $u$ is given by $u(t) = M\sum_{i\in\mathbb{Z}_{\geq 0}} \mathbb{1}_{(t_{2i},t_{2i+1})}$, where $(t_i)_{i\in\mathbb{Z}_{\geq 0}}$ is an increasing sequence of times in $[0, T]$. We denote by $p$ the corresponding solution to (10.4).

  We want to compare with the control $\bar{u} = M\mathbb{1}_{[T-C/M,T]}$, for which the corresponding solution to (10.4) is denoted $\bar{p}$. Then, $\bar{p}(T) = G_M^{-1}(C/M)$.

  Let us show that $p(T) < \bar{p}(T) = G_M^{-1}(C/M)$. We use an induction to prove that for all $i \in \mathbb{Z}_{\geq 0}$, $p(t_{2i}) < G_M^{-1}(C/M)$. Indeed, if we assume that for a $i \in \mathbb{Z}_{\geq 0}^*$, we have for every $k \leq i$, $p(t_{2k}) < G_M^{-1}(C/M) < \theta$. Then, on $(t_{2i}, t_{2i+1})$, we solve the equation

  $$\dot{p} = f(p), \qquad p(t_{2i}) < \theta.$$

  Since $f < 0$ on $(0, \theta)$, it implies that $p$ is decreasing on $(t_{2i}, t_{2i+1})$, thus $p(t_{2i+1}) < p(t_{2i})$. On $[t_{2i+1}, t_{2i+2})$, we have

  $$G_M(p(t_{2i+2})) = G_M(p(t_{2i+1})) + t_{2i+2} - t_{2i+1} < G_M(p(t_{2i})) + t_{2i+2} - t_{2i+1}.$$

  By induction, we deduce that

  $$\begin{aligned} G_M(p(t_{2i+2})) <& G_M(p(t_{2i-2})) + t_{2i} - t_{2i-1} + t_{2i+2} - t_{2i+1} \\ <& G(p(t_0)) + t_2 - t_1 + \ldots + t_{2i+2} - t_{2i+1} \leq C/M, \end{aligned}$$

  since $p(t_0) = 0$ and $\sum_{k=0}^{i}(t_{2k+2} - t_{2k+1}) \leq \frac{C}{M}$. We infer

  $$p(t_{2i+2}) < G_M^{-1}(C/M).$$

  This concludes the induction and the proof in this first case.

- *Case* $p_M(C/M) > \theta$. We use the same strategy and introduce the solution $p$ to (10.4) with $u$ given by $u(t) = M\sum_{i\in\mathbb{Z}_{\geq 0}} \mathbb{1}_{(t_{2i},t_{2i+1})}$, where $(t_i)_{i\in\mathbb{Z}_{\geq 0}}$ is an increasing sequence of time in $[0, T]$. We want to compare with the solution $\bar{p}$ for $\bar{u} = M\mathbb{1}_{[0,C/M]}$.

We first observe that since $\bar{p}(C/M) = p_M(C/M) > \theta$ and $f > 0$ on $(\theta, 1)$, we have $\bar{p}$ increasing on $[C/M, T]$ and $\bar{p}(C/M) = G_M^{-1}(C/M)$. If at time $t_1$, we have $p(t_1) < \theta$, then on $(t_1, t_2)$, $p$ is decreasing. Then $p(t_2) < p(t_1) \leq \bar{p}(t_1 - t_0)$ and we may prove as above that as long as $p(t_{2i+1}) < \theta$, we have $p(t_{2i+2}) < \bar{p}\left(\sum_{k=0}^{i}(t_{2k+1} - t_{2k})\right)$.

As a consequence the solution $p$ associated with the optimal control should satisfy $p(t_1) > \theta$. Then, on $(t_1, T)$ the function $p$ solving (10.4) is increasing, thus on $(t_1, T)$, we have $p > \theta > p^*$. Then the switch function $w$ is increasing. However, we have $w > \Lambda$ on $(t_1, t_2)$ since $u = 0$ from Lemma 10.6. Hence, it is not possible to have $u = M$ for larger times.

- *In the case where* $p_M(C/M) = \theta$. In this case, we have $u = M\mathbb{1}_{(\lambda, C/M+\lambda)}$ for any $0 \leq \lambda \leq T - C/M$. Indeed, for such a function, we have $p \equiv 0$ on $[0, \lambda]$ and $p \equiv \theta$ on $[C/M + \tau, T]$. By contradiction, assume there is an interval on which $u = 0$ between two intervals on which $u = M$, then on this interval $p$ is decreasing, and thus $p$ cannot reach the value $\theta$ at the final time of control, by comparison.

□

### 10.A.5 Proof of Corollary 10.1

According to Proposition 10.2, we know that $(u^\epsilon)_{\epsilon>0}$ converges weak star in $L^\infty(0,T)$ to a solution of Problem ($\mathcal{P}_{\text{reduced}}$), say $u^*$.

Since $u^*$ is an extremal point of the convex set $\mathcal{U}_{T,C,M}$ to which all elements of the sequence $(u^\epsilon)_\epsilon$ belong, it follows from [23] that the $L^\infty$-weak$*$ convergence (that is here, $L^1$-weak convergence) implies strong convergence in $L^1$, and therefore

$$\lim_{\epsilon \to 0} \int_0^T |u^\epsilon(t) - u^*(t)| dt = 0.$$

Finally, we conclude by observing that, whenever $C \neq C^*(M)$, the solution to Problem ($\mathcal{P}_{\text{reduced}}$) is unique according to Theorem 10.1.

## 10.B Qualitative properties of the minimizers

In this section, we treat a slight extension of ($\mathcal{P}_{\text{full}}$) to more general final time criteria. We state in Proposition 10.3 useful qualitative properties of solutions, under additional assumptions on the biological parameters.

We let $G : \mathbb{R}_+^2 \to \mathbb{R}$ be a smooth function such that $\partial_1 G \geq 0 \geq \partial_2 G$, and define $\mathcal{J}(u) = G(\mathbf{n}(T))$.

**Proposition 10.3.** *Let $u^*$ be a local minimizer of $\mathcal{J}$ in $\mathcal{U}_{T,C,M}$ On the set $I_* := \{u^* \in (0, M)\}$ we have $u^* = \big(H - f_2\big)(\mathbf{n})$, where*

$$H = \frac{-\partial_2 f_1 \partial_2 f_2 \partial_1 f_1 + \partial_1 f_2 (\partial_2 f_1)^2 - f_1 \partial_{21}^2 f_2 \partial_2 f_1 + \partial_2 f_2 f_1 \partial_{21}^2 f_1}{\partial_{22}^2 f_2 \partial_2 f_1 - \partial_{22}^2 f_1 \partial_2 f_2}.$$

*Let $\overline{N}(C) := K \max(1 - \frac{d_1}{b_1}, 1 - \frac{d_2}{b_2}) + C$,*

$$N_{\min} := K \min_{p \in [0,1]} \frac{\big(b_1(1 - s_h p) - d_1\big)(1 - p) + \big(b_2 - d_2\big)p}{b_1(1-p)(1 - s_h p) + b_2 p}.$$

*and*

$$M(C) := \max_{(N,p) \in [N_{\min}, \overline{N}(C)] \times [0,1]} \big(H((1-p)N, pN) - f_2((1-p)N, pN)\big).$$

*Then $M(C)$ is finite. In addition if $M > M(C)$ and the biological parameters satisfy a nonlinear condition $\mathcal{G} \geq 0$ (defined below in (10.24)) then there exists $t_0, t_1 \in [0, T]$ with $t_0 M + (T - t_1)M \leq C$ such that set $I_M := \{u^* = M\}$ is equal to $[0, t_0] \cup [t_1, T]$.*

**Interpretation.** Under reasonable assumptions on the biological parameters, the optimal replacement strategies use the maximal possible release flux only at the beginning and at the end of the protocol.

We split the proof into

- estimates on the dynamics of (10.2)-(10.3) yielding bounds on $H$ and other quantities,
- the use of a first-order necessary optimality condition.

### 10.B.1 Estimates on the dynamics

We normalize (10.2)-(10.3) by letting $x = n_1/K$, $y = n_2/K$, so by an abuse of notations

$$\begin{cases} f_1(x,y) = b_1 x(1 - s_h y/(x+y))(1-(x+y)) - d_1 x, \\ f_2(x,y) = b_2 y(1-(x+y)) - d_2 y. \end{cases}$$

We compute:

$$\partial_1 f_1(x,y) = b_1(1-(2x+y))(1-s_h\tfrac{y}{x+y}) - d_1 + s_h b_1(1-(x+y))\tfrac{xy}{(x+y)^2},$$

$$\partial_2 f_1(x,y) = -b_1 x(1-s_h\tfrac{y}{x+y}) - s_h b_1(1-(x+y))\tfrac{x^2}{(x+y)^2},$$

$$\partial_1 f_2(x,y) = -b_2 y,$$

$$\partial_2 f_2(x,y) = b_2(1-(x+2y)) - d_2.$$

We also compute the second-order derivatives of the form $\partial^2_{2i} f_j$, pour $i,j \in \{1,2\}$:

$$\partial^2_{21} f_1(x,y) = -b_1(1-s_h\tfrac{y}{x+y}) + s_h b_1 \tfrac{x^2-xy}{(x+y)^2} - 2s_h b_1 \tfrac{xy}{(x+y)^3}(1-(x+y)),$$

$$\partial^2_{22} f_1(x,y) = 2s_h b_1 \tfrac{x^2}{(x+y)^3},$$

$$\partial^2_{21} f_2(x,y) = -b_2,$$

$$\partial^2_{22} f_2(x,y) = -2b_2.$$

The function $\chi := \partial_2 f_2(\mathbf{n})\partial^2_{22} f_1(\mathbf{n}) - \partial_2 f_1(\mathbf{n})\partial^2_{22} f_2(\mathbf{n})$ involved in the denominator of $H$ reads

$$\chi(x,y) = 2s_h b_1 \frac{x^2}{(x+y)^3}(1-(x+y))\big(b_2(1-(x+2y)) - d_2\big) \\ - 2b_2\big(b_1 x(1-s_h\frac{y}{x+y}) - s_h b_1(1-(x+y))\frac{x^2}{(x+y)^2}\big).$$

In general we have

$$\chi(x,y) = 2b_1 b_2 x\Big(\frac{s_h x}{(x+y)^3}\big(1-(x+y)\big)\big(1-d_2/b_2-(2x+3y)\big) - (1-s_h\frac{y}{x+y})\Big).$$

With an abuse of notations and letting $s_2 = d_2/b_2 \in (0,1))$ we rewrite $\chi$ as

$$\chi(p,N) = 2b_1 b_2(1-p)\Big(\frac{s_h(1-p)(1-s_2)}{N} - s_h(1-p)(2+p) - N(1-s_h)\Big),$$

where as usual $p = y/N$ and $N = x+y$.

**Remark 10.3.** *In the special case when there is no cytoplasmic incompatibility, $s_h = 0$ and*

$$\chi = -2b_2 b_1 x < 0.$$

In all generality, the sign of $\chi$ is given by

$$Q(p,N) := s_h(1-s_2)(1-p) - s_h(1-p)(2+p)N - N^2(1-s_h),$$

which we need to compute along trajectories.

**Lemma 10.9.** *Let $N_+$ be the following, non-negative and bounded function on $[0,1]$:*

$$N_+(p) := \frac{-s_h(1-p)(2+p) + \sqrt{s_h^2(1-p)^2(2+p)^2 + 4s_h(1-s_h)(1-s_2)(1-p)}}{2(1-s_h)}.$$

*Then $\chi(p,N) < 0$ with $N > 0$ if and only if $N > N_+(p)$.*

*Proof.* We compute the roots of $Q$ as a polynomial in $N$, $Q(p, N_\pm(p)) = 0$ where

$$N_\pm(p) = \frac{-s_h(1-p)(2+p) \pm \sqrt{s_h^2(1-p)^2(2+p)^2 + 4s_h(1-s_h)(1-s_2)(1-p)}}{2(1-s_h)}.$$

Obviously, $N_-(p) \le 0 \le N_+(p)$. Since the leading coefficient of $Q$ is negative, it follows that $Q(p,N) < 0$ if and only if $N > N_+(p)$ or $N < N_-(p)$. The sign of $Q$ gives the sign of $\chi$, and the result is proved. □

For a solution to (10.2)-(10.3) we have

$$\frac{dN}{dt} = N\Big(\big(b_1(1-s_hp) - d_1\big)(1-p) + \big(b_2 - d_2\big)p - \big(b_1(1-p)(1-s_hp) + b_2p\big)N\Big) + u.$$

Since $u \ge 0$, in particular $N$ is increasing as soon as

$$N \le \frac{\big(b_1(1-s_hp) - d_1\big)(1-p) + \big(b_2 - d_2\big)p}{b_1(1-p)(1-s_hp) + b_2p} =: \psi(p).$$

Let $N_{\min} := \min_{p\in[0,1]} \psi(p)$ and $\overline{N}_+ := \max_{p\in[0,1]} N_+(p)$.

**Lemma 10.10.** *The following inequality holds: $N_{\min} \ge \overline{N}_+$ if and only if $\mathcal{G}(\delta, s_1, s_2, s_h) \ge 0$ (for some non-linear function $\mathcal{G}$ defined below in (10.24)).*

*In this case, $\chi < 0$ along all trajectories of (10.2)-(10.3) with $u \ge 0$, $n_1(0) \ge 0$, $n_2(0) \ge 0$ and $n_1(0) + n_2(0) > N_{\min}$.*

*Proof.* First, by the above computations we have that $\frac{dN}{dt} > 0$ as soon as $N < \psi(p)$. If $N < N_{\min}$ then this always holds, and in particular the set $\{N \ge N_{\min}\}$ is absorbing for the dynamics of (10.2)-(10.3) with $u \ge 0$. Therefore, $N_{\min} \ge \overline{N}_+$ suffices to have that $N > \overline{N}_+$ along all the trajectories in consideration, and thus $\chi < 0$ by Lemma 10.9.

Secondly we notice that $N_+(1) = 0$, and we compute, letting $\lambda = (1-s_h)(1-s_2)/s_h$:

$$2\frac{1-s_h}{s_h}N_+'(p) = 1 + 2p - \frac{(1-p)(2+p)(1+2p) + 2\lambda}{\sqrt{(1-p)^2(2+p)^2 + 4\lambda(1-p)}}.$$

In particular, $\lim_{1^-} N_+' = -\infty$.

In addition $N_+' = 0$ if and only if

$$(1+2p)^2\big((1-p)^2(2+p)^2 + 4\lambda(1-p)\big) = \big((1-p)(2+p)(1+2p) + 2\lambda\big)^2,$$

which is equivalent to

$$-(1+2p)(1-p)^2 = \lambda.$$

There is no solution $p \in [0,1]$ since $1 + 2p > 0$ and $\lambda > 0$, so $N_+$ is monotone decreasing and therefore

$$\overline{N}_+ = N_+(0) = \frac{s_h}{1-s_h}\Big(-1 + \sqrt{1 + \frac{(1-s_h)(1-s_2)}{s_h}}\Big).$$

(Note that as $s_h \to 1$, $\overline{N}_+ \to (1-s_2)/2$.)

Thirdly, we rewrite $\delta = d_2/d_1$ and

$$\psi(p) = 1 - s_1\frac{(\delta-1)p + 1}{(1-p)(1-s_hp) + \frac{\delta s_1}{s_2}p}.$$

A direct computation shows that $\psi'$ is positively proportional to

$$R(p) := s_h(\delta-1)p^2 + 2s_hp + \delta\big(\frac{s_1}{s_2} - 1\big) - s_h,$$

which is a second-order polynomial in $p$. We compute its discriminant

$$\Delta_R = 4s_h\delta\big(s_h - (\delta-1)(\frac{s_1}{s_2} - 1)\big).$$

Let $\xi = \frac{(\delta-1)(\frac{s_1}{s_2}-1)}{s_h}$. The roots of $R$ if $\Delta_R \geq 0$ (that is $\xi \leq 1$) are given by

$$p_\pm^R = \frac{-1 \pm \sqrt{\delta(1-\xi)}}{\delta - 1}.$$

We gather the various cases below, using the notations $p_{\min} = \arg\min \psi$:

- If $\delta < 1$, then either $\xi < 1$ and $p_-^R > p_+^R > 0$ with $p_-^R > 1$, so the minimum of $\psi$ is reached at $p_{\min} = \min(1, p_+^R)$, or $\xi \geq 1$ and $\psi$ is decreasing, in which case $p_{\min} = 1$.
- If $\delta = 1$ then $R(p) = 2s_h p + \frac{s_1}{s_2} - 1 - s_h$. Let $\zeta = \frac{1+s_h-\frac{s_1}{s_2}}{2s_h}$. If $\zeta \leq 0$ then the minimum of $\psi$ is reached at $p_{\min} = 0$, if $\zeta \geq 1$ then the minimum of $\psi$ is reached at $p_{\min} = 1$ and if $\zeta \in (0,1)$ then the minimum of $\psi$ is reached at $p_{\min} = \zeta$. In short, $p_{\min} = \min(\max(0,\zeta),1)$.
- If $\delta > 1$:
  - if $\xi \geq 1$ then $p_{\min} = 0$ ($\psi$ is increasing);
  - if $\xi < 1$:
    - if $\delta(1-\xi) \leq 1$ then the two roots of $R$ are negative and $p_{\min} = 0$;
    - if $\delta(1-\xi) > 1$ then $\psi$ is either decreasing or decreasing-increasing, and $p_{\min} = \min(1, p_+^R)$;

We can summarize all this into:

$$p_{\min} = \min\big(\max\big(0, p_m(\delta, \frac{s_1}{s_2}, s_h)\big), 1\big),$$

where $p_m$ is the following continuous map $\mathbb{R}_+ \times \mathbb{R}_+ \times (0,1] \to \mathbb{R}$:

$$p_m(\delta,\sigma,s_h) = \begin{cases} \frac{-1+\sqrt{\delta\left(1-\frac{(\delta-1)(\sigma-1)}{s_h}\right)}}{\delta-1} \text{ if } \delta \neq 1 \text{ and } (\delta-1)(\sigma-1) < s_h, \\ \frac{-1}{\delta-1} \text{ if } \delta \neq 1 \text{ and } (\delta-1)(\sigma-1) \geq s_h, \\ \frac{1+s_h-\sigma}{2s_h} \text{ if } \delta = 1. \end{cases}$$

Back to $\psi$, we find that $N_{\min}$ can be equal to either $\psi(0) = 1 - s_1$, $\psi(1) = 1 - s_2$, $\psi(\frac{1+s_h-\sigma}{2s_h}) = 1 - \frac{4s_1 s_h}{(1+s_h-\sigma)(3-\sigma-s_h)}$ (in the case $\delta = 1$ and $0 < \frac{1+s_h-\sigma}{2s_h} < 1$), or to

$$\psi\Big(\frac{-1+\sqrt{\delta\big(1-\frac{(\delta-1)(\sigma-1)}{s_h}\big)}}{\delta-1}\Big) = 1 - \frac{s_1(\delta-1)^2\sqrt{\delta\big(1-\frac{(\delta-1)(\sigma-1)}{s_h}\big)}}{\delta^2\big(s_h-(\sigma-1)(\delta-1)\big) + \big((\delta-1)(\sigma\delta-1) - s_h(1+\delta)\big)\sqrt{\delta\big(1-\frac{(\delta-1)(\sigma-1)}{s_h}\big)}}$$

if $\delta \neq 1$, $(\delta-1)(\sigma-1) < s_h$ and $0 < \frac{-1+\sqrt{\delta\left(1-\frac{(\delta-1)(\sigma-1)}{s_h}\right)}}{\delta-1} < 1$.

Finally, we obtain that $N_{\min} \geq \overline{N}_+$ if and only if $\mathcal{G} \geq 0$, where

$$\boxed{\mathcal{G}(\delta,s_1,s_2,s_h) = \psi(p_{\min}) - \frac{s_h}{1-s_h}\big(-1+\sqrt{1+\frac{(1-s_h)(1-s_2)}{s_h}}\big),} \tag{10.24}$$

where we recall that

$$\psi(p) = 1 - s_1\frac{(\delta-1)p+1}{(1-p)(1-s_h p)+\delta\sigma p}, \quad p_{\min} = \min\big(\max\big(0, p_m(\delta,\frac{s_1}{s_2},s_h)\big),1\big),$$

and

$$p_m(\delta,\sigma,s_h) = \begin{cases} \frac{-1+\sqrt{\delta\left(1-\frac{(\delta-1)(\sigma-1)}{s_h}\right)}}{\delta-1} \text{ if } \delta \neq 1 \text{ and } (\delta-1)(\sigma-1) < s_h, \\ \frac{-1}{\delta-1} \text{ if } \delta \neq 1 \text{ and } (\delta-1)(\sigma-1) \geq s_h, \\ \frac{1+s_h-\sigma}{2s_h} \text{ if } \delta = 1, \end{cases}$$

and where the second term in $\mathcal{G}$ must be replaced by its limit $s_h \to 1$ if $s_h = 1$, namely by $(1-s_2)/2$. □

**Remark 10.4.** *For instance, if $s_h = 1$ and $\delta = 1$ we find that*

$$\mathcal{G}(1, s_1, s_2, 1) = \frac{1 + s_2 - s_1/s_2}{2},$$

*so that the property holds if and only if $s_1 \leq s_2(1+s_2)$. The usual set of assumptions for Wolbachia is $\delta \geq 1$ (it is life-shortening) and $s_1 \leq s_2$ (the wild population is fitter than the introduced one, recalling that the population at equilibrium is $K(1 - s_i)$, where $K$ is the environmental carrying capacity).*

*The property holds in particular if CI is perfect ($s_h = 1$), there is no life-shortening effect ($\delta = 1$) and the wild population has higher birth rate, $b_1 \geq b_2$ (but of course it holds for many other cases as well).*

**Remark 10.5.** *We can also notice that since $-1 + \sqrt{1+z} \leq z/2$ for all $z \geq 0$, we have that if $\psi(p) \geq (1 - s_2)/2$ for all $p \in [0, 1]$ then $\mathcal{G} \geq 0$. By taking the even more restrictive sufficient condition $\psi(p) \geq (1 - s_2)/2$ for all $p \in \mathbb{R}$, we get the following, simpler sufficient condition:*

$$\big(\delta\frac{s_1}{2}(\frac{1}{s_2} - 1) + s_1 - \frac{(1 + s_h)(1 + s_2)}{2}\big)^2 \leq s_h(1 + s_2 - 2s_1).$$

*However, the condition $\mathcal{G} \geq 0$ is obviously better than this one.*

By a straightforward computation using the equation on $N$, we can bound uniformly the total population along trajectories of (10.2)-(10.3):

**Lemma 10.11.** *If $N(0) \leq \max_{p\in[0,1]} \psi(p)$ then along all trajectories with $u \geq 0$ and $\int_0^T u(t)dt \leq C$ we have*

$$N \leq \max_{p\in[0,1]} \psi(p) + C =: \overline{N}(C).$$

*In addition,*

$$\overline{N}(C) = \max(1 - s_1, 1 - s_2) + C. \tag{10.25}$$

*Proof.* Equation (10.25) follows from the previous study of $\psi$, in the proof of Lemma 10.10. □

Finally, we bound the function $H$ (defining the value of the optimal controls along singular arcs in Proposition 10.3):

**Lemma 10.12.** *Assume $\mathcal{G} \geq 0$. Then the function $H : [N_{\min}, \overline{N}(C)] \times [0, 1] \to \mathbb{R}$ from (5.8) is bounded.*

*Proof.* It appears that the denominator in (5.8) is $\chi$, which is negative and vanishes only at $p = 1$, with order 1 if $s_h < 1$, and order 2 if $s_h = 1$. (Indeed, by Lemma 10.10, $\mathcal{G} \geq 0$ is sufficient to get this property for $N \geq N_{\min}$). To get the boundedness of $H$, we must therefore check that the numerator also vanishes at $p = 1$, with appropriate order in $(1 - p)$.

The numerator is comprised of four terms, being equal to

$$-\partial_2 f_1 \partial_2 f_2 \partial_1 f_1 + \partial_1 f_2 (\partial_2 f_1)^2 - f_1 \partial_{21}^2 f_2 \partial_2 f_1 + \partial_2 f_2 f_1 \partial_{21}^2 f_1.$$

The first three terms have $\partial_2 f_1$, which is equal to

$$-b_1(1 - p)(1 - s_h p)N - s_h b_1(1 - N)(1 - p)^2,$$

and is therefore of order 1 if $s_h < 1$, and of order 2 if $s_h = 1$.

The last term has

$$\partial_{21}^2 f_1 = b_1\Big( - (1 - s_h p) + s_h(1 - p)^2 + s_h p(1 - p) - \frac{2s_h p(1 - p)}{N}\Big),$$

which is of order 0 is $s_h < 1$, and of order 1 if $s_h = 1$, and it also has $f_1$, which is alwyas of order 1. Therefore the numerator of $H$ is globally of order 1 in $(1 - p)$ at $p = 1$ if $s_h < 1$, and of order 2 if $s_h = 1$, and thus $H$ is bounded. □

### 10.B.2 Derivation of qualitative properties of minimizers

To conclude the proof of Proposition 10.3 we simply need to use the first-order necessary condition of optimality (as in Lemma 10.6).

Indeed, a straightforward computation shows that $u^*$ (associated with a solution $\mathbf{n}^*$ to (10.2)-(10.3)) is a local minimizer only if there exists $\Lambda < 0$ such that

- $I_M := \{u^* = M\} = \{q_2 < \Lambda\}$,
- $I_* := \{u^* \in (0, M)\} = \{q_2 = \Lambda\}$,
- $I_0 := \{u^* = 0\} = \{q_2 > \Lambda\}$,

where $\mathbf{q} = (q_1, q_2)$ solves

$$-\frac{d\mathbf{q}}{dt} = \begin{pmatrix} \partial_1 f_1 & \partial_1 f_2 \\ \partial_2 f_1 & \partial_2 f_2 \end{pmatrix} \mathbf{q}, \quad \mathbf{q}(T) = \begin{pmatrix} \partial_1 G(\mathbf{n}^*(T)) \\ \partial_2 G(\mathbf{n}^*(T)) \end{pmatrix}$$

The fact that $\Lambda < 0$ comes from $q_2(t) < 0$ on $[0, T)$, which is a direct consequence of the assumptions on $G$ ($\partial_1 G \geq 0 \geq \partial_2 G$) and the fact that the system is competitive: $\partial_1 f_2, \partial_2 f_1 < 0$.

From this, we can derive the expression of $H$: by $q_2 \equiv \Lambda$ on $I_*$ we can deduce that for $t \in \mathring{I}_*$,

$$\partial_2 f_1(\mathbf{n}(t)) q_1(t) = -\partial_2 f_2(\mathbf{n}(t))\Lambda.$$

This equality holds on a neighborhood of $t$, and we can use it in the equation on $q_1$ to deduce that

$$-\frac{dq_1}{dt} = \Lambda \frac{d}{dt}\Big(\frac{-\partial_2 f_2(\mathbf{n}(t))}{\partial_2 f_1(\mathbf{n}(t))}\Big) = -\Lambda \partial_1 f_1(\mathbf{n}(t)) \frac{\partial_2 f_2(\mathbf{n}(t))}{\partial_2 f_1(\mathbf{n}(t))} + \Lambda \partial_1 f_2(\mathbf{n}(t)).$$

After division by $\Lambda$, and recalling that $\frac{d\mathbf{n}}{dt} = \mathbf{f}(\mathbf{n}) + \begin{pmatrix} 0 \\ u \end{pmatrix}$ we obtain the expression of $H$.

It only remains to prove the claim that $I_M = [0, t_0] \cup [t_1, T]$. It can be seen easily that the switch function $q_2$ solves in fact the following Cauchy problem:

$$\begin{cases} a(\mathbf{n})\ddot{q}_2 + b(\mathbf{n}, u^*)\dot{q}_2 + c(\mathbf{n}, u^*)q_2 = 0 \text{ on } [0, T], \\ q_2(T) = -\big(K(1 - \frac{d_2}{b_2}) - n_2(T)\big)_+, \\ \dot{q}_2(T) = -\partial_2 f_1(\mathbf{n}(T)) n_1(T) + \partial_2 f_2(\mathbf{n}(T))(X_2 - n_2(T))_+, \end{cases} \tag{10.26}$$

where

$$c(\mathbf{n}, u) = \chi(\mathbf{n})(u + f_2(\mathbf{n}) - H(\mathbf{n})), \quad a(\mathbf{n}) = -\partial_2 f_1(\mathbf{n}),$$
$$b(\mathbf{n}, u) = -(\partial_1 f_1(\mathbf{n}) + \partial_2 f_2(\mathbf{n}))\partial_2 f_1(\mathbf{n}) + f_1(\mathbf{n})\partial^2_{12} f_1(\mathbf{n}) + (f_2(\mathbf{n}) + u)\partial^2_{22} f_1(\mathbf{n}).$$

Thus, as long as $\chi < 0$ on a trajectory associated with an optimal control $u^*$, and $M > \max(f_2 - H)$ then $q_2$ has no local minimum in $I_M$. Since $q_2 < \Lambda$ in the interior of $I_M$ and $q_2 = \Lambda$ at boundary points of $I_M$ that are interior to $[0, T]$, this suffices to show that $I_M = [0, t_0] \cup [t_1, T]$. The conditions are met, as can be seen from the facts gathered in Section 10.B.1: if the parameters satisfy $\mathcal{G} \geq 0$ then $\chi < 0$ on all trajectories (Lemma 10.9), and if $M > M(C)$ then $M > \max(f_2 - H)$ where the maximum is taken on the admissible set.

# Chapter 11

# Sharp seasonal threshold property for cooperative population dynamics with concave nonlinearities

This chapter is a work in collaboration with Hongjun Ji.

**Abstract.** We consider a biological population whose environment varies periodically in time, exhibiting two very different "seasons": one is favorable and the other one is unfavorable. For monotone differential models with concave nonlinearities, we address the following question: the system's period being fixed, under what conditions does there exist a critical duration for the unfavorable season? By "critical duration" we mean that above some threshold, the population cannot sustain and extincts, while below this threshold, the system converges to a unique periodic and positive solution. We term this a "sharp seasonal threshold property" (SSTP, for short). Building upon a previous result, we obtain sufficient conditions for SSTP in any dimension and apply our criterion to a two-dimensional model featuring juvenile and adult populations of insects.

## 11.1 Introduction

We study differential dynamical systems arising from nonlinear periodic positive differential equations of the form

$$\frac{dx}{dt} = F(t,x), \tag{11.1}$$

where $F$ is monotone and concave. These systems exhibit well-known contraction properties when $F$ is continuous (see [113], [207], [126]). We extend in Theorem 11.1 these properties to nonlinearities that are only piecewise-continuous in time. This extension is motivated by the study of typical seasonal systems in population dynamics.

We denote by $\theta \in [0,1]$ the proportion of the year spent in unfavorable season. Then, we convene that time $t$ belongs to an unfavorable (resp. a favorable) season if $nT \leq t < (n+\theta)T$ (resp. if $(n+\theta)T \leq t < (n+1)T$) for some $n \in \mathbb{Z}_+$. In other words, we study the solutions to:

$$\frac{dX}{dt} = G(\pi_\theta(t), X), \quad \pi_\theta(t) = \begin{cases} \pi^U \text{ if } \frac{t}{T} - \lfloor \frac{t}{T} \rfloor \in [0,\theta), \\ \pi^F \text{ if } \frac{t}{T} - \lfloor \frac{t}{T} \rfloor \in [\theta,1), \end{cases} \tag{11.2}$$

for some $G : \mathcal{P} \times \mathbb{R}^N \to \mathbb{R}^N$, with $\pi^U, \pi^F \in \mathcal{P}$ where $\mathcal{P}$ is the parameter space. We are looking for conditions ensuring that a sharp seasonal threshold property holds, that is:

$$\exists \theta_* \in [0,1] \text{ such that } \begin{cases} \text{if } \theta < \theta_*, \exists! q : \mathbb{R}_+ \to \mathbb{R}^N, T\text{-periodic}, q \gg 0 \text{ and} \\ \forall X_0 \in \mathbb{R}^N_+ \backslash \{0\}, X \text{ converges to } q, \\ \text{if } \theta > \theta_*, \forall X_0 \in \mathbb{R}^N_+, \ X \text{ converges to } 0. \end{cases} \tag{SSTP}$$

Ecologically, the respective duration of dry and wet seasons is crucial for population sustainability in various species. The property (SSTP) means that if the dry season is longer than $\theta_* T$ then the population collapses and if it is shorter then the population densities will tend to be periodic.

Assume that $F(t,0) \equiv 0$. Thanks to the contraction properties of concave nonlinearities, the whole problem reduces to the study of the Floquet eigenvalue with maximum modulus of the linearization of (11.1) at $X = 0$:

$$\frac{dz}{dt} = D_x F(t,0) z. \tag{11.3}$$

In fact, this eigenvalue is equal to the spectral radius of the Poincaré application for (11.3), which we compute here for piecewise-autonomous systems.

Our proof uses the Perron-Frobenius theorem and relies on the Perron eigenvalue and (left and right) eigenvectors. The importance of this eigenvalue for quantifying the effects of seasonality has been acknowledged continuously in mathematical biology in at least three application fields: circadian rhythms (in particular in connection with cell division and tumor growth), harvesting and epidemiology.

It was noted in [56] that Floquet eigenvalue with maximum modulus of (11.3) is always larger that the Perron eigenvalue of some averaged (over a period) matrix $\overline{F}$ defined from the entries of $D_x F(t,0)$. There has been a continued interest in this eigenvalue for linear models of cell division since and we refer to [95] in particular for a detailed study of the monotonicity of the Perron eigenvalue with respect to parameters of a structured model for cell division. In a stochastic framework for growth and fragmentation, [43] establishes a similar monotonicity property. In this context, the Perron eigenvalue is seen as the cell growth rate, and this is why its dependence in the model parameters is important. Here, we connect the eigenvalue monotonicity with a non-extinction condition to derive the (SSTP). We emphasize that our Theorem 11.2 gives some sufficient conditions for the monotonicity of the Perron eigenvalue, in the case when there are only two different seasons.

In dimension 1, for the logistic equation with harvesting, Xiao has shown in [236] a sharp threshold property, where the two different "seasons" correspond to one harvesting period ("unfavorable season") and one rest period ("favorable season"). Contrary to the case of cell division, the model treated there is non-linear, though 1-dimensional. Our results extend a part of those of [236] to $n$-dimensional concave monotone systems. Note that the cited article also studies the maximal sustainable yield, which can be seen as an objective function of the periodic solution $q$. On this topic, [170, Section 5] studies a structured problem of adaptive dynamics with concave nonlinearity and periodic forcing to show a similar effect as in [236] (there, for population size): in both cases, time fluctuations can improve an objective value.

For applications in epidemiology, where seasonality often has dramatic effects, we refer to [20] and [19] for the computation of case reproduction numbers with seasonal forcing.

The organization of the paper is as follows. The motivating model is detailed in Section 11.2, where we also define some concepts. In Section 11.3 we state our results: first (Theorem 11.1) an extension to piecewise-continuous nonlinearities of the well-known results on monotone concave nonlinearities, then (Theorem 11.2) fairly general sufficient conditions for systems in any space dimension $N \in \mathbb{Z}_{>0}$ to satisfy (SSTP), and finally (Theorem 11.3) an application to the two-dimensional system (11.2), for which we are able to show the threshold property (SSTP) for a wide range of parameters. The proofs are detailed in Section 11.4 (and in Appendix 11.A for Theorem 11.1), while extensions and possible research directions are gathered in Section 11.5.

## 11.2 Context and motivation

Our reference model is a simplistic description of the population dynamics of some insects, with a juvenile stage exposed to quadratic competition and an adult stage. Let $J(t), A(t)$ represent the populations of juveniles and adults at time $t$, respectively. A very simple dynamic is defined by

$$\begin{cases} \dfrac{dJ}{dt} = bA - J(h + d_J + c_J J), \\ \dfrac{dA}{dt} = hJ - d_A A, \end{cases} \tag{11.4}$$

where $d_Y$ ($Y \in \{J, A\}$) stands for the (linear) death rate, $b$ is the birth rate, $h$ is the hatching rate and the parameter $c_J$ tunes the only non-linearity: quadratic competition (=density-dependent

death rate) among juveniles. This term effectively limits the total population size, as we will prove below. We use it to represent resource limitation both for breeding sites availability and for nutrient availability during growth. In principle, the parameters may depend on time:

$$\forall t \in \mathbb{R}, \quad \pi(t) := (b, h, d_J, c_J, d_A) \in \mathbb{R}^5_+. \tag{11.5}$$

For convenience, we rewrite the right-hand side of (11.4) as $G(\pi, X)$ with $X = (J, A) \in \mathbb{R}^2$, and $G : \mathbb{R}^5_+ \times \mathbb{R}^2 \to \mathbb{R}^2$.

In the tempered areas where mosquito populations are established, dramatic seasonal variations in population abundance are usually observed. Namely, there is explosive growth in summer after rain events, whereas mosquitoes are very scarce in winter. This phenomenon is possible thanks to dormant (or "quiescent" or "refuge") phases in the mosquito's life-cycle. These seasonal variations imply that the natural environment (temperature, rainfall, humidity etc.) is very important for the mosquito.

We propose to study population dynamics in simple models such as (11.4) under periodic seasonal forcing. As a rough approximation, we set up (11.4) with periodic piecewise-constant coefficients of period $T = 1$ year, each one possibly taking two different values over one period. Thus, the year is divided into unfavorable and favorable seasons, defined by parameter values $\pi^U, \pi^F \in \mathbb{R}^5_+$ such that

$$\begin{pmatrix} -d_J^F + d_J^U & b^F - d_A^F - (b^U - d_A^U) \\ h^F - h^U & -d_A^F + d_A^U \end{pmatrix} > 0. \tag{11.6}$$

The four scalar inequalities of condition (11.6) deserve a biological justification. It implies that during the favorable season, the hatching rate is larger than during the unfavorable season, while death rates (for juveniles, and adults) are smaller. These assumptions rely on the facts that breeding sites availability and quality is much higher in good season (whence higher hatching rate and birth rate and lower juvenile competition), while the temperature increase can be expected to extend the life-span of both adults and juveniles. The first component in (11.6) implies that the growth coefficients $b - d_A$ are ordered: $b^F - d_A^F > b^U - d_A^U$. This is true in particular if $b^F > b^U$, but holds in more generality.

We emphasize that the systems under study are excessively simple because, in mathematical terms, they are cooperative with concave nonlinearity, and as such they have strong asymptotic convergence properties.

Let $F : \mathbb{R}_t \times \mathbb{R}^N_x \to \mathbb{R}^N$ be piecewise continuous in $t$ and continuously differentiable in $x$. The system (11.1) is *cooperative* if its Jacobian matrix is Metzler:

$$\forall (t, x) \in \mathbb{R}_+ \times \mathbb{R}^N_+, \, i \neq j \implies \frac{\partial F_i}{\partial x_j}(t, x) \geq 0, \tag{M}$$

It is *positive* (*i.e.*, $\mathbb{R}^N_+$ is an invariant set) if

$$\forall t \in \mathbb{R}_+, \, \forall 1 \leq i \leq N, \, \forall x \geq 0, \quad x_i = 0 \implies F_i(t, x) \geq 0. \tag{P}$$

Under condition (M), (11.1) is positive if $\forall t \in \mathbb{R}_+$, $F(t, 0) \geq 0$. We say that (11.1) defines a *concave dynamics* on $\mathbb{R}^N_+$ if

$$\forall 0 \ll x \ll y, \, D_x F(t, x) \geq D_x F(t, y), \tag{C}$$

and that (11.3) is *irreducible* if

$$\forall t \in \mathbb{R}_+, \, D_x F(t, 0) \text{ is irreducible in } M_N(\mathbb{R}). \tag{I}$$

## 11.3 Results

### 11.3.1 General results

In order to study the asymptotic behavior of (11.2), we generalize a result by Smith [207] (refined by Jiang in [126]) about continuous concave and cooperative nonlinearities to piecewise-continuous (in time) nonlinearities.

**Theorem 11.1.** *Let $F : \mathbb{R}_t \times \mathbb{R}^N_x \to \mathbb{R}^N$ be $T$-periodic and piecewise-continuous in $t$ and such that for all $t \in \mathbb{R}_+$, $F(t, \cdot) \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}^N)$. Assume that $F$ satisfies assumptions (P), (M), (C) and (I), so that the associated differential system (11.1) is positive, monotone and concave with irreducible linearization at 0. Let $\lambda \in \mathbb{R}$ denote the Floquet multiplier with maximal modulus of (11.3).*

*If $\lambda \leq 1$ then every non-negative solution of (11.1) converges to 0. Otherwise,*

*(i) either every non-negative solution of (11.1) satisfies $\lim_{t\to\infty} x(t) = \infty$,*

*(ii) or (11.1) possesses a unique (nonzero) $T$-periodic solution $q(t)$.*

*In case (ii), $q \gg 0$ and $\lim_{t\to\infty}(x(t) - q(t)) = 0$ for every non-negative solution of (11.1).*

The proof of Theorem 11.1 (in Appendix 11.A) follows closely the lines of [207] and [126].

An illuminating example when Theorem 11.1 applies is for $T$-periodic piecewise autonomous differential systems, where for all $x \in \mathbb{R}^N$, $F(\cdot, x)$ is a piecewise-constant function. Namely, we assume that there exists $K \in \mathbb{Z}_{>0}$ and functions $(F^k)_{1\leq k\leq K} : \mathbb{R}^N_+ \to \mathbb{R}^N_+$ such that:

$$F(t, x) = F^k(x) \text{ if } \frac{t}{T} - \lfloor \frac{t}{T} \rfloor \in [\theta_{k-1}, \theta_k), \tag{11.7}$$

where $(\theta_k)_{0\leq k\leq K} \in [0, 1]^{K+1}$ is a non-decreasing family such that $\theta_0 = 0$ and $\theta_K = 1$. To verify the hypotheses of Theorem 11.1, we need to assume that for all $1 \leq k \leq K$, $F^k$ is continuously differentiable, monotone, concave and satisfies $F^k(0) = 0$; and in addition that $DF^k(0)$ is irreducible for all $1 \leq k \leq K$.

The main advantage of piecewise-constant non-linearities is that for such dynamics (and almost only for these dynamics), the Floquet multiplier with maximal modulus $\lambda$ can be computed explicitly as the following spectral radius:

$$\lambda = \rho\big(e^{(\theta_K - \theta_{K-1})T \cdot DF^K(0)} \cdots e^{(\theta_1 - \theta_0)T \cdot DF^1(0)}\big). \tag{11.8}$$

In the case $K = 2$, with $\theta := \theta_1$, the Perron-Frobenius theorem applies to

$$M(\theta) := e^{(1-\theta)T \cdot DF^2(0)} e^{\theta T \cdot DF^1(0)},$$

which is positive since $DF^k(0)$ are (irreducible) Metzler matrix by (M) (and (I)). Therefore there exists unique vectors $V(\theta), V_*(\theta) \gg 0$ with $\|V(\theta)\| = 1$ and $\langle V(\theta), V_*(\theta)\rangle = 1$, and a unique positive number $\rho(\theta)$ such that

$$M(\theta)V(\theta) = \rho(\theta)V(\theta), \quad M(\theta)^* V_*(\theta) = \rho(\theta)V_*(\theta). \tag{11.9}$$

In this setting, assume without loss of generality that $\mu(DF^2(0)) \geq \mu(DF^1(0))$, and denote $S := DF^1(0) - DF^2(0)$. We consider two specific cases:

(A) $DF^1(0)$ and $DF^2(0)$ have the same principal right or left eigenvector;

(B) for all $\theta \in [0, 1]$, one of the following holds:

(B-1) $\exists P \in GL_N(\mathbb{R})$, $PS < 0$ and $(P^{-1})^* V_*(\theta) > 0$;

(B-2) $\exists P \in GL_N(\mathbb{R})$, $SP < 0$ and $P^{-1}V(\theta) > 0$;

(B-3) $\exists P, Q \in M_N(\mathbb{R})$, $S < P^*Q$ and $PV_*(\theta) = -QV(\theta)$.

**Theorem 11.2.** *Let $F$ of the form (11.7) with $K = 2$ satisfy the assumptions of Theorem 11.1. Assume that the forward orbits of (11.1) are bounded. Then under (A) or (B), (SSTP) holds.*

**Remark 11.1.** *In addition, condition $(B-1)$ (resp. $(B-2)$) is equivalent to*

$$S^* V_*(\theta) < 0 \text{ (resp. } SV(\theta) < 0),$$

*and if condition (A) holds then $V(\theta) \equiv V$ or $V_*(\theta) \equiv V_*$, where $V$ (resp. $V_*$) is the right (resp. left) principal eigenvector of $DF^i(0)$, $i \in \{1, 2\}$.*

*Proof.* We apply Theorem 11.1 and check that the value of $\lambda$ (determining if case $(i)$ or $(ii)$ occurs) is a decreasing function of $\theta$ under assumptions (A) or (B). The forward-boundedness of orbits rules out the case $x \to +\infty$, thus leading to the result. More details in Section 11.4.1. □

**Remark 11.2.** *In the case $DF^2(0) > DF^1(0)$, we note that conditions (B-1) and (B-2) are obviously satisfied with $P = I$ (identity matrix), and condition (B-3) is obviously satisfied with $P = Q = 0$.*

**Remark 11.3.** *As will be seen below, in practical situations it is sometimes easier to check condition (B-1) rather than computing $S^*V_*(\theta)$.*

### 11.3.2 Application to a two-dimensional model of insect population dynamics

We can now specify Theorem 11.2 to the two-dimensional ($N = 2$) case of (11.4). First we describe the general properties of this system

**Proposition 11.1.** *For system (11.4) written as $\dot{X} = G(\pi(t), X) =: F(t, X)$, where $\pi$ is defined by (11.5), assume that $\pi(t) \gg 0$, there exists $c, C \in \mathbb{R}_+^*$ such that $\pi_i(t) \geq c$ for $i \in \{4, 5\}$ and $\pi(t) \leq C\mathbb{1}$. Then, it is positive, forward-bounded, cooperative and concave.*

Then, we give the dynamics of the non-seasonal (=autonomous) system (11.4) with $\pi(t) \equiv \pi = (b, h, d_J, c_J, d_A)$. We define the basic offspring number:

$$\mathcal{R}_0 = \mathcal{R}(\pi) := \frac{bh}{d_A(h + d_J)}. \tag{11.10}$$

**Proposition 11.2.** *If $\mathcal{R}_0 \leq 1$, then (11.4) has no positive steady state and the trivial equilibrium is a global attractor. If $\mathcal{R}_0 > 1$ then (11.4) has exactly one positive steady state $S_1^* = (\mathcal{R}_0 - 1)\left(\frac{h+d_J}{c_J}, \frac{h(h+d_J)}{c_J d_A}\right)$, which is a global attractor in $\mathbb{R}_+^2 \backslash \{0\}$.*

The proofs of Proposition 11.2 and Proposition 11.1 are to be found in Section 11.4.2.

We finally state the sharp seasonal threshold property for (11.2):

**Theorem 11.3.** *For (11.2) under assumption (11.6), if $\mathcal{R}_0(\pi^U) < 1 < \mathcal{R}_0(\pi^F)$ and $b^U + d_J^U > d_A^U$ (where $\pi^U = (b^U, h^U, d_J^U, c_J^U, d_A^U)$) then (SSTP) holds with $\theta_* \in (0, 1)$.*

*Proof.* We check assumption $(B - 1)$ with

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad (P^{-1})^* = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

More details in Section 11.4.3. □

**Remark 11.4.** *If instead of (11.6) we assume the stronger condition*

$$\begin{pmatrix} -(h^F + d_J^F) + h^U + d_J^U & b^F - b^U \\ h^F - h^U & -d_A^F + d_A^U \end{pmatrix} > 0, \tag{11.11}$$

*then assumption $(B - 1)$ (or $(B - 2)$) of Theorem 11.2 applies with $P = I$ and no further computations are needed.*

*We emphasize that (11.6) is more biologically relevant than (11.11). The latter requires that the increase of the hatching rate between favorable and unfavorable season does more than compensate the decrease of juvenile death rate, which is highly debatable. This justifies the technical computations of Section 11.4.3.*

*Note that in any case, no assumptions are made on $c_J^U$ and $c_J^F$, since the behavior is only determined by the linearization at 0.*

## 11.4 Proofs

### 11.4.1 Proof of Theorem 11.2

When there are only two dynamics within a period, that is when $K = 2$, we notice that the alternative $(i) - (ii)$ from Theorem 11.1 is uniquely determined by the sign of the real function:

$$\theta \quad \mapsto \quad \rho\left(e^{(1-\theta)T \cdot DF^2(0)} e^{\theta T \cdot DF^1(0)}\right) - 1.$$

We notice that

**Lemma 11.1.** *The function $\rho : [0,1] \to \mathbb{R}$ is $\mathcal{C}^1$ and satisfies*

$$\rho'(\theta) = T\rho(\theta)\langle (DF^1(0) - DF^2(0))V(\theta), V_*(\theta)\rangle. \tag{11.12}$$

*Proof.* By Perron-Frobenius theorem, $\rho(\theta)$ is the maximal root of the characteristic polynomial of $M(\theta)$, whose entries are analytic functions of $\theta$. In particular, it is $\mathcal{C}^1$.

The principal eigenvector of norm 1 of $M(\theta)$, that is $V(\theta)$, depends smoothly of $\theta$, as can be seen by uniqueness for all $\theta$. Then, $V_*(\theta)$ also depends smoothly of $\theta$ since the same argument applies to $M^*(\theta)$ and $V_*(\theta)$ is equal to the principal eigenvector $Y_*(\theta)$ of $M^*(\theta)$ divided by $\langle V(\theta), Y_*(\theta)\rangle > 0$, which is a smooth function of $\theta$.

Let us write $M_i := DF^i(0)$ for $i \in \{1,2\}$. We differentiate the identity $\rho(\theta) = \langle M(\theta)V(\theta), V_*(\theta)\rangle$ to obtain

$$\begin{aligned}
\rho'(\theta) &= \langle M(\theta)V'(\theta), V_*(\theta)\rangle + \langle M'(\theta)V(\theta), V_*(\theta)\rangle + \langle M(\theta)V(\theta), V_*'(\theta)\rangle,\\
&= \rho(\theta)\Big(\langle V'(\theta), V_*(\theta)\rangle + T\big(\langle V(\theta), M_1^*V_*(\theta)\rangle - \langle M_2V(\theta), V_*(\theta)\rangle\big) + \langle V(\theta), V_*'(\theta)\rangle\Big),\\
&= T\rho(\theta)\langle (M_1 - M_2)V(\theta), V_*(\theta)\rangle,
\end{aligned}$$

since $M'(\theta) = Te^{(1-\theta)TM_2}\big(M_1 - M_2\big)e^{\theta TM_1}$ and $\langle V(\theta), V_*(\theta)\rangle \equiv 1$. □

Applying Theorem 11.1 with the assumption that the forward orbits are bounded, we are left with either global asymptotic stability of 0 is $\lambda \leq 1$, or the global stability of the unique positive periodic solution, if $\lambda > 1$. Using formula (11.8), we obtain (SSTP) with $\rho(\theta_*) = 1$ (or $\theta_* = 0$ if $\rho(0) > 1$, and $\theta_* = 1$ if $\rho(1) \leq 1$) if $\rho$ is a decreasing function of $\theta$.

It remains to prove that any of the conditions $(A)$ or $(B)$ implies that $\rho$ is decreasing. Under assumption $(B-1)$, with $S = DF^1(0) - DF^2(0)$ we get by Lemma 11.1

$$\frac{\rho'(\theta)}{T\rho(\theta)} = \langle SV(\theta), V_*(\theta)\rangle = \langle PSV(\theta), (P^{-1})^*V_*(\theta)\rangle < 0,$$

since $PS < 0$, $V(\theta) \gg 0$ and $(P^{-1})^*V_*(\theta) > 0$ by assumption. Note that this condition is equivalent to $S^*V_*(\theta) < 0$. Reasoning by density of $GL_N(\mathbb{R})$ in $M_N(\mathbb{R})$, we assume that $S$ is invertible and check that if $S^*V_* < 0$ then $P = -S^{-1}$ satisfies the assumption, and conversely if $PS = Q < 0$, upon writing $(P^{-1})^* = (Q^{-1})^*S^*$ we get $(Q^{-1})^*S^*V_* > 0$, and by multiplication by $Q^* < 0$ this implies $S^*V_* < 0$. The argument is symmetrical for assumption $(B-2)$ and is omitted here.

Under assumption $(B-3)$ we get by Lemma 11.1

$$\frac{\rho'(\theta)}{T\rho(\theta)} = \langle SV(\theta), V_*(\theta)\rangle < \langle P_*(\theta)Q(\theta)V(\theta), V_*(\theta)\rangle = -\|Q(\theta)V(\theta)\|^2 \leq 0,$$

since $V(\theta), V_*(\theta) \gg 0$ (for the inequality), and $PV_* = -QV$ (for the equality).

Finally, under assumption $(A)$ we get that $V(\theta) \equiv V$ and $V_*(\theta) \equiv V_*$ where $V$ (resp. $V_*$) is the principal eigenvector (resp. left principal eigenvector) of $DF^1(0)$ (which is the same as the one of $DF^2(0)$). In this case,

$$\frac{\rho'(\theta)}{T\rho(\theta)} = \langle SV, V_*\rangle = \mu(DF^1(0)) - \mu(DF^2(0)),$$

whence the result.

### 11.4.2 Proofs of Proposition 11.1 and Proposition 11.2

Recall that by definition,

$$\forall X \in \mathbb{R}^2, \quad F(t,X) = G(\pi(t), X) := \begin{pmatrix} \pi_1 X_2 - (\pi_2 + \pi_3 + \pi_4 X_1)X_1 \\ \pi_2 X_1 - \pi_5 X_2 \end{pmatrix}.$$

We first proceed to the proof of Proposition 11.1. If $X_i = 0$ for some $i \in \{1,2\}$, then since $\pi(t) \geq 0$, $F_i(t,X) \geq 0$. Therefore the system is positive.

We recall the notation $\pi = (b, h, d_J, c_J, d_A)$. We have:

$$D_X F = \begin{pmatrix} -h - d_J - 2c_J J & b \\ h & -d_A \end{pmatrix}.$$

Thus, $D_X F$ is a Metzler matrix, so (11.4) is monotone cooperative.

To check the concavity property, let $X \gg Y$. We simply compute

$$D_X F(t, X) - D_X F(t, Y) = \begin{pmatrix} 2c_J(Y_1 - X_1) & 0 \\ 0 & 0 \end{pmatrix} > 0.$$

Then, we proceed to the proof of Proposition 11.2. Calculating the equations of nullclines

$$\begin{aligned} bA - hJ - d_J J - c_J J^2 &= 0, \\ hJ - d_A A &= 0, \end{aligned}$$

immediately yields all steady states as:

$$S_0^* = (0, 0), \quad S_1^* = (\frac{bh}{d_J} - h - d_J)(\frac{1}{c_J}, \frac{h}{c_J d_A}).$$

Then, the sign of both components of $S_1^*$ is equal to the sign of $\mathcal{R}_0 - 1$, whence the result.

The stability and local behavior of solutions is detailed in

**Proposition 11.3.** *If $\mathcal{R}_0 \leq 1$ the unique equilibrium point $S_0^* = (0, 0)$ is either a stable node (when $\mathcal{R}_0 < 1$) or a singular point of superior order and of attracting type (when $\mathcal{R}_0 = 1$), in which case all the orbits in the neighborhood of the $S_0^*$ tend to $S_0^*$ along direction $\theta_1 := \arctan \frac{h+d_J}{b}$.*

*If $\mathcal{R}_0 > 1$, the equilibrium point $S_0^* = (0, 0)$ is of saddle type, and the direction of unstable manifold is $\frac{h+d_J-d_A+\sqrt{(h+d_J-d_A)^2+4bh}}{2b}$. The equilibrium point $S_1^*$ is a stable node.*

*Proof.* We divide the proof into three parts, depending on the sign of $\mathcal{R}_0 - 1$.

**When $\mathcal{R}_0 = 1$.** Then (11.4) becomes

$$\begin{aligned} \frac{dJ}{dt} &= -\frac{bh}{d_A} J + bA - c_J J^2, \\ \frac{dA}{dt} &= hA - d_A A. \end{aligned} \tag{11.13}$$

The determinant of its Jacobian matrix is

$$\begin{vmatrix} -\frac{bh}{d_A} & b \\ h & -d_A \end{vmatrix} = 0.$$

Hence, the equilibrium point $S_0^*$ of system (11.13) is an isolated critical point of higher order.

Obviously, system (11.13) is analytic in a neighborhood of the origin. By [241, Theorem 3.10, p. 70], any orbit of (11.13) tending to the origin must tend to it spirally or along a fixed direction, which depends on the characteristic equation of system (11.13). First of all, we introduce the polar coordinates $J = r\cos\delta$, $A = r\sin\delta$, where $\delta \in [0, \frac{\pi}{2}]$, $r \in \mathbb{R}_+$ and we get the relation

$$\begin{cases} \dot{r} = r^{-1}(J\dot{J} + A\dot{A}) = r^m[R(\delta) + o(1)], \\ \dot{\delta} = r^{-2}(J\dot{A} - A\dot{J}) = r^{m-1}[G(\delta) + o(1)]. \end{cases}$$

This yields

$$\begin{cases} \dot{r} = r(-\frac{bh}{d_A}\cos^2\delta + b\cos\delta\sin\delta + h\cos\delta\sin\delta - d_A\sin^2\delta - c_J r\cos^3\delta), \\ \dot{\delta} = h\cos^2\delta - d_A\cos\delta\sin\delta + (h + d_J)\cos\delta\sin\delta - b\sin^2\delta + c_J r\cos^2\delta\sin\delta. \end{cases}$$

Then the characteristic equation of system (11.13) takes the form

$$G(\delta) = h\cos^2\delta - d_A\cos\delta\sin\delta + (h + d_J)\cos\delta\sin\delta - b\sin^2\delta = 0, \tag{11.14}$$

and we have

$$R(\delta) = -\frac{bh}{d_A}\cos^2\delta + b\cos\delta\sin\delta + h\cos\delta\sin\delta - d_A\sin^2\delta.$$

After equation (11.14), we get

$$(\frac{h+d_J}{b}\cos\delta - \sin\delta)(d_A\cos\delta + b\sin\delta) = 0. \tag{11.15}$$

Thus

$$\begin{cases}\tan\ \delta_1 = \frac{h+d_J}{b},\\ \tan\delta_2 = -\frac{d_A}{b}.\end{cases}$$

Clearly, $G(\delta) = 0$ has two real roots which we denote by $\delta_1$ and $\delta_2$. By the results in [241, Section 2], we know that neither the case no orbit of system (11.13) can tend to the critical point $S_0^*$ spirally nor the singular case (if $G(\delta) \equiv 0$).

The orbits of the system tend to the origin along a characteristic direction $\delta_i$, given by solutions of the equation (11.14). Since the system is positive we need to consider $\delta \in [0, \frac{\pi}{2}]$, so $\delta_1 = \arctan\frac{h+d_J}{b}$ is in first orthant and the orbits of the system approach the origin along the direction $\delta = \delta_J$.

**When $\mathcal{R}_0 > 1$.** We now write the Jacobian matrix **Jac** of the system

$$\mathbf{Jac} := \begin{pmatrix}-h - d_J - 2c_J E & b\\ h & -d_A\end{pmatrix},$$

and consider $\mathbf{Jac}_0$ and $\mathbf{Jac}_1$ are the Jacobian matrices respectively at equilibrium point $S_0^*$ and $S_1^*$. At ${S_0}^*$,

$$\mathbf{Jac}_0 = \begin{pmatrix}-h - d_J & b\\ h & -d_J\end{pmatrix},$$

whose eigenvalues read

$$\lambda_1 = \frac{-(h+d_J+d_A)+\sqrt{\Delta}}{2},$$
$$\lambda_2 = \frac{-(h+d_J+d_A)-\sqrt{\Delta}}{2},$$

where $\Delta := (h + d_J + d_A)^2 - 4[(h + d_J)d_A - hb] > 0$ (since $(h + d_J)d_A - hb < 0$). Then

$$\lambda_1 + \lambda_2 = -(h + d_J + d_A) < 0,$$
$$\lambda_1\lambda_2 = (h + d_J)d_A - hb < 0,$$

so that one eigenvalue is positive and the another one is negative: $S_0^*$ is a saddle point.

To find the direction of the stable manifold or unstable manifold at $S_0^*$, we write

$$\frac{\dot{A}}{\dot{J}} = \frac{dA}{dt} = \frac{hJ - d_A A}{-hJ - d_J J + bA - c_J J^2} = \frac{h - \frac{A}{J}}{-h - d_J + b\frac{A}{J} - c_J J}.$$

Consider $(J, A)$ tending to $S_0^*$ and let $k := \frac{A}{J}$. Then $k$ is a solution to

$$k = \frac{h - d_A k}{-h - d_J + bk},$$

which leads to two solutions $(k_1, k_2) \in \mathbb{R}_+^* \times \mathbb{R}_-^*$ given by

$$\frac{h + d_J - d_A \pm \sqrt{(h + d_J - d_A)^2 + 4bh}}{2b}.$$

Hence, the boundary lines are $A = k_1 J$ and $A = k_2 J$ and by unstable manifold theorem we know that $k_1$ is the direction of unstable manifold at $(0, 0)$

Then, at equilibrium point $S_1^*$,

$$\mathbf{Jac}_1 = \begin{pmatrix}h + d_J - \frac{2bh}{d_A} & b\\ h & -d_A\end{pmatrix},$$

whose eigenvalues $\lambda_1, \lambda_2$ are real and satisfy

$$\lambda_1 + \lambda_2 = h + d_J - \tfrac{2bh}{d_A} - d_A < 0,$$
$$\lambda_1\lambda_2 = -d_A(h + d_J) + bh > 0.$$

This implies that the two eigenvalues are real and negative, hence $S_1^*$ is a stable node.

**Finally, if $\mathcal{R}_0 < 1$.** Then at equilibrium point $S_0^*$

$$\mathbf{Jac}_0 = \begin{pmatrix} -h - d_J & b \\ h & -d_A \end{pmatrix}.$$

Because $(h + d_J)d_A - hb > 0$, the eigenvalues are such that

$$\begin{aligned} \lambda_1 + \lambda_2 &= -(h + d_J + d_A) < 0, \\ \lambda_1\lambda_2 &= (h + d_J)d_A - hb > 0, \end{aligned}$$

with also the discriminant $(-h - d_J + d_A)^2 + 4bh > 0$, hence they are both negative and the equilibrium point $S_0^*$ is a stable node. □

**Remark 11.5.** *In particular when $h = 0$ (no hatching), and the trivial equilibrium point $S_0^*$ is a stable node.*

We now prove that all the orbits of (11.4) are forward bounded.

**Lemma 11.2.** *Let*

$$\tau^* := \sup_{t\geq 0} \frac{h(t)}{d_A(t)}, \quad J^* := \sup_{t\geq 0} \frac{b(t)\tau^* - h(t) - d_J(t)}{c_J(t)}.$$

*Under the assumptions of Proposition 11.1, $\tau^*$ and $J^*$ are finite. For all $X_0 \in \mathbb{R}^2_+$ and all real number $L \geq \max(0, J^*)$ such that $X_0 \in \Omega_L := [0, L] \times [0, \tau^* L]$, the solution $X(t)$ of (11.4) with initial data $X_0$ belongs to $\Omega_M$.*

*Proof.* Under the assumptions of Proposition 11.1, $c_J \geq c > 0$ and $d_A \geq c$ while all parameters are smaller than $C > 0$, hence $J^*$ and $\rho^*$ are finite.

For $L > 0$ we define the area rectangle $\Omega_L$ surrounded by four line segments $\ell_i$ with outward normal vector $\nu_i$:

$$\begin{aligned} \ell_1 &= \{(J, A) | J = 0, 0 \leq A \leq \tau^* L)\}, \quad \nu_1 = (-1, 0), \\ \ell_2 &= \{(J, A) | J = L, 0 \leq A \leq \tau^* L)\}, \quad \nu_2 = (1, 0), \\ \ell_3 &= \{(J, A) | 0 \leq J \leq L, A = 0\}, \quad \nu_3 = (0, -1) \\ \ell_4 &= \{(J, A) | 0 \leq J \leq L, A = \tau^* L\}, \quad \nu_4 = (0, 1). \end{aligned}$$

To prove that $\Omega_L$ is positively invariant, since the system is positive, we only need to show that the scalar products of $\frac{dX}{dt}$ and $\nu_i$ on $\ell_i$ for $i \in \{2, 4\}$ are non-positive:

$$\begin{aligned} \nu_4 \cdot G(\pi, X) &= hJ - d_A\tau^* L \leq 0 \text{ since } J \leq L \text{ and } d_A\tau^* \geq h, \\ \nu_2 \cdot G(\pi, X) &= bA - hL - d_J L - c_J L^2. \end{aligned}$$

Since $A < \tau^* L$, $\nu_2 \cdot G(\pi, X) \leq 0$ on $\ell_2$ as soon as $b\tau^* - h - d_J - c_J L \leq 0$, that is

$$L \geq \frac{b\tau^* - h - d_J}{c_J}.$$

Upon taking $L \geq J^*$ this inequality is satisfied. For $L$ large enough such that $X_0 \in \Omega_L$, we have proved that for all $t > 0$, the solution $X(t)$ of (11.4) belongs to $\Omega_L$. □

The Dulac (divergence) criterion ensures that the system has no limit cycle, since:

$$\mathrm{div}(F) = -(h + d_J + c_J J + d_A) < 0.$$

This concludes the proof.

### 11.4.3 Proof of Theorem 11.3

Theorem 11.3 is a consequence of Theorem 11.2, condition (B-1). To check this condition, we apply the following result (specific to the dimension $N = 2$) to the positive matrix $M(\theta)$:

**Lemma 11.3.** *Let $S \in M_2(\mathbb{R})$ be a positive matrix, and assume vector $W = (w_1, w_2) \gg 0$ satisfies $S^*W = \mu W$ for some $\mu > 0$ (i.e. $W$ is the principal eigenvector of $S^*$). Then, $w_2 > w_1$ if and only if*

$$s_{11} + s_{21} < s_{12} + s_{22}, \tag{11.16}$$

*Where $s_{11}$, $s_{21}$, $s_{12}$ and $s_{22}$ are the elements of matrix $S$.*

*Proof.* We write $SW = \mu W$ as

$$\begin{cases} s_{11}w_1 + s_{21}w_2 = \mu w_1, \\ s_{12}w_1 + s_{22}w_2 = \mu w_2, \end{cases} \iff \begin{cases} s_{11} + s_{21}\frac{w_2}{w_1} = \mu, \\ s_{12}\frac{w_1}{w_2} + s_{22} = \mu. \end{cases}$$

If $0 < w_1 < w_2$, since $S \gg 0$ we deduce that $s_{11} + s_{21} < \rho < s_{12} + s_{22}$.
Conversely, if $s_{11} + s_{21} < s_{12} + s_{22}$, subtracting the previous equalities we obtain

$$\mu(1 - \frac{w_2}{w_1}) = s_{11} - s_{12} + \frac{w_2}{w_1}(s_{21} - s_{22}) < (s_{22} - s_{21})(1 - \frac{w_2}{w_1}).$$

By contradiction, we assume that $w_2 < w_1$. Then $\mu < s_{22} - s_{21}$. Injecting this inequality into the previous equality we obtain

$$s_{12} + \frac{w_2}{w_1}s_{22} < (s_{22} - s_{21})\frac{w_2}{w_1},$$

whence $s_{12} < -\frac{w_2}{w_1}s_{21}$, which contradicts $S > 0$. Hence $w_2 > w_1$. □

Lemma 11.3 is satisfied by $M(\theta)$, so that condition $(B-1)$ holds with $P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Indeed, $(P^{-1})^* = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ and $(P^{-1})^* V_* > 0$ with $V_* \gg 0$ if and only if $[V_*]_2 > [V_*]_1$, hence by (11.6) we have $P\big(DF^2(0) - DF^1(0)\big) < 0$.

The remaining of the proof is devoted to checking that $M_{12}(\theta) + M_{22}(\theta) - M_{11}(\theta) - M_{21}(\theta) > 0$. To this aim, we diagonalize

$$DF^1(0) = \begin{pmatrix} -h^U - d_J^U & b^U \\ h^U & -d_A^U \end{pmatrix} \text{ and } DF^2(0) = \begin{pmatrix} -h^F - d_J^F & b^F \\ h^F & -d_A^F \end{pmatrix}$$

by

$$DF^1(0) = P_U \begin{pmatrix} \lambda_U^+ & 0 \\ 0 & \lambda_U^- \end{pmatrix} P_U^{-1}, \quad DF^2(0) = P_F \begin{pmatrix} \lambda_F^+ & 0 \\ 0 & \lambda_F^- \end{pmatrix} P_F^{-1},$$

where for $\sharp \in \{U, F\}$,

$$P_\sharp = \begin{pmatrix} 1 & 1 \\ x_\sharp^+ & x_\sharp^- \end{pmatrix}, \quad P_\sharp^{-1} = \frac{1}{x_\sharp^- - x_\sharp^+} \begin{pmatrix} x_\sharp^- & -1 \\ -x_\sharp^+ & 1 \end{pmatrix}$$

and

$$\begin{aligned} \lambda_\sharp^\pm &= -\frac{1}{2}(h^\sharp + d_J^\sharp + d_A^\sharp) \pm \frac{1}{2}\sqrt{(h^\sharp + d_J^\sharp - d_A^\sharp)^2 + 4h^\sharp b^\sharp}, \\ x_\sharp^\pm &= \frac{\lambda_\sharp^\pm + h^\sharp + d_J^\sharp}{b^\sharp}, \\ &= \frac{1}{2b^\sharp}(h^\sharp + d_J^\sharp - d_A^\sharp) \pm \frac{1}{2b^\sharp}\sqrt{(h^\sharp + d_J^\sharp - d_A^\sharp)^2 + 4h^\sharp b^\sharp}. \end{aligned}$$

The condition of Lemma 11.3 will follow from:

**Lemma 11.4.** *For $\sharp \in \{U, F\}$, we have $x_\sharp^- < 0 < x_\sharp^+$ and $1 + x_\sharp^- > 0$.*

*Proof.* The first inequalities follow directly from the above expression of $x_\sharp^\pm$. Then, we compute $1 + x_\sharp^- = \frac{2b^\sharp + h^\sharp + d_J^\sharp - d_A^\sharp - \sqrt{(h^\sharp + d_J^\sharp - d_A^\sharp)^2 + 4h^\sharp b^\sharp}}{2b^\sharp}$. We have

$$\begin{aligned} (2b^\sharp + h^\sharp + d_J^\sharp - d_A^\sharp)^2 &= 4(b^\sharp)^2 + 4b^\sharp(h^\sharp + d_J^\sharp - d_A^\sharp) + (h^\sharp + d_J^\sharp - d_A^\sharp)^2 \\ &> (h^\sharp + d_J^\sharp - d_A^\sharp)^2 + 4h^\sharp b^\sharp \end{aligned}$$

since $b^\sharp + d_J^\sharp - d_A^\sharp > 0$ (explicit assumption in Proposition 11.2 for $\sharp = U$, and from $\mathcal{R}(\pi^F) > 1$ for $\sharp = F$). It implies $1 + x_\sharp^- > 0$.

□

Thanks to the above diagonalization, we can write $M = M(\theta) = (m_{ij})_{1\le i,j\le 2}$ as

$$\begin{aligned}
m_{11} &= (\beta^+ x_F^- - \beta^- x_F^+)(\gamma^+ x_U^- - \gamma^- x_U^+) + (-\beta^+ + \beta^-)(x_U^+ x_U^- \gamma^+ - x_U^+ x_U^- \gamma^-),\\
m_{12} &= (\beta^+ x_F^- - \beta^- x_F^+)(-\gamma^+ + \gamma^-) + (-\beta^+ + \beta^-)(-x_U^+ \gamma^+ + x_U^- \gamma^-),\\
m_{21} &= (x_F^+ x_F^- \beta^+ - x_F^+ x_F^- \beta^-)(\gamma^+ x_U^- - \gamma^- x_U^+) + (-x_F^+ \beta^+ + x_F^- \beta^-)(x_U^+ x_U^- \gamma^+ - x_U^+ x_U^- \gamma^-),\\
m_{22} &= (x_F^+ x_F^- \beta^+ - x_F^+ x_F^- \beta^-)(-\gamma^+ + \gamma^-) + (-x_F^+ \beta^+ + x_F^- \beta^-)(-x_U^+ \gamma^+ + x_U^- \gamma^-),
\end{aligned}$$

where

$$\begin{aligned}
&\beta^+ := e^{\lambda_F^+(1-\theta)T}, \quad \beta^- := e^{\lambda_F^-(1-\theta)T},\\
&\gamma^+ := e^{\lambda_U^+\theta T}, \quad \gamma^- := e^{\lambda_U^-\theta T},\\
&\alpha := \frac{b^U b^F}{\sqrt{\big((h^U + d_J^U - d_A^U)^2 + 4h^U b^U\big)\big((h^F + d_J^F - d_A^F)^2 + 4h^F b^F\big)}}.
\end{aligned}$$

Proving $m_{11} + m_{21} < m_{12} + m_{22}$ therefore amounts to checking

$$\begin{aligned}
&\beta^+\gamma^+(x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + \beta^+\gamma^-(x_U^- - x_F^-)(1 + x_F^+)(1 + x_U^+)\\
&\quad + \beta^-\gamma^+(x_U^+ - x_F^+)(1 + x_U^-)(1 + x_F^-) + \beta^-\gamma^-(x_F^+ - x_U^-)(1 + x_F^-)(1 + x_U^+) < 0. \quad (11.17)
\end{aligned}$$

We introduce $\Psi : \mathbb{R}_+^2 \to \mathbb{R}$ as

$$\begin{aligned}
\Psi(\beta,\gamma) := {}& \beta\gamma(x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + \beta(x_U^- - x_F^-)(1 + x_F^+)(1 + x_U^+)\\
&+ \gamma(x_U^+ - x_F^+)(1 + x_U^-)(1 + x_F^-) + (x_F^+ - x_U^-)(1 + x_F^-)(1 + x_U^+),
\end{aligned}$$

so that (11.17) is equivalent to $\Psi(\frac{\beta^+}{\beta^-}, \frac{\gamma^+}{\gamma^-}) < 0$. First, it is easily checked that $\Psi(1,1) = 0$, $\beta^+ > \beta^-$ and $\gamma^+ > \gamma^-$. Then, by Lemma 11.4, $x_F^- < 0 < x_U^+$ and $1 + x_\sharp^\flat > 0$ for $\sharp \in \{U, F\}$ and $\flat \in \{+, -\}$. Hence for $\beta > 1$, we have

$$\begin{aligned}
\frac{\partial\Psi(\beta,\gamma)}{\partial\gamma} &= \beta(x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + (x_U^+ - x_F^+)(1 + x_U^-)(1 + x_F^-)\\
&< (x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + (x_U^+ - x_F^+)(1 + x_U^-)(1 + x_F^-)\\
&= (x_F^- - x_F^+)(1 + x_U^-)(1 + x_U^+).
\end{aligned}$$

Symmetrically, for $\gamma > 1$ we have

$$\begin{aligned}
\frac{\partial\Psi(\beta,\gamma)}{\partial\beta} &= \gamma(x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + (x_U^- - x_F^-)(1 + x_F^+)(1 + x_U^+)\\
&< (x_F^- - x_U^+)(1 + x_F^+)(1 + x_U^-) + (x_U^- - x_F^-)(1 + x_F^+)(1 + x_U^+)\\
&= (x_U^- - x_U^+)(1 + x_F^-)(1 + x_F^+).
\end{aligned}$$

Applying Lemma 11.4 again, we deduce that if $\beta, \gamma > 1$ then

$$\frac{\partial\Psi}{\partial\gamma}, \frac{\partial\Psi}{\partial\beta} < 0.$$

In particular $\Psi(\frac{\beta^+}{\beta^-}, \frac{\gamma^+}{\gamma^-}) < 0$, and this concludes the proof.

## 11.5 Discussion and extensions

**Geometric viewpoint.** We denote by $\Upsilon \times \Upsilon_*$ the graph of $\upsilon := (V, V_*) : [0,1] \to (\mathbb{R}_+^*)^{2N}$. Then we define $r(\theta) := \frac{\rho'(\theta)}{T\rho(\theta)} = \langle SV(\theta), V_*(\theta)\rangle$. Denoting by $\psi_S : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ the bilinear form $(V, W) \mapsto \langle AV, W\rangle$, we get $r = \psi_S \circ \upsilon$. Let $X_S := \{\psi_S < 0\}$, it is an open and radial subset of $\mathbb{R}^{2N}$ (if $Y \in X_S$ and $\lambda > 0$, then $\lambda Y \in X_S$). $\rho(M)$ is decreasing if and only if $r$ is decreasing, which is equivalent to $\Upsilon \times \Upsilon_* \subset X_S$. Up to changing $S$ into $-S$, assumption (11.18) amounts to $\upsilon(0), \upsilon(1) \in X_S$.

The case (A) implies that $\Upsilon \times \Upsilon_*$ is a singleton. Then (11.18) simply rewrites $(\mu_2 - \mu_1)^2 > 0$.

**Practical computations in higher dimension.** Theorem 11.2 suggests 4 different sufficient conditions on $DF^1(0)$ and $DF^2(0)$ to obtain (SSTP). Apart from the trivial situations when $DF^1(0) - DF^2(0)$ has a sign or when the two matrices share the same principal eigenvector, how applicable are these conditions when $N > 2$ If $DF^i(0)$ is diagonalizable for $i \in \{1, 2\}$, which we write

$$DF^i(0) = P_i^{-1}\mathbf{diag}((\lambda_i^{(k)})_{1\le k\le N})P_i,$$

then we can compute

$$M_{i,j}(\theta) = \sum_{j',j''=1}^{N} P_1^{-1}(i,j')Q(j',j'')P_2(j'',j)e^{T\left(\theta\lambda_1^{(j')}+(1-\theta)\lambda_2^{(j'')}\right)},$$

where $Q(j',j'') = \sum_{k=1}^N P_1(j',k)P_2^{-1}(k,j'')$. For any matrix $\Gamma = (\gamma(i,j))_{1\le i,j\le N} \in GL_N(\mathbb{R})$ such that $\Gamma M(\theta) > 0$, we obtain $\Gamma V(\theta) > 0$ (where $V(\theta)$ is the principal eigenvector of $M(\theta)$). Then, a sufficient condition for (SSTP) is given by $(DF^2(0) - DF^1(0))\Gamma^{-1} < 0$. Symmetrically, if $M(\theta)\Gamma > 0$ then a sufficient condition is given by $\Gamma^{-1}(DF^2(0) - DF^1(0)) < 0$.

In order to get better conditions than the obvious ones, we require that $\Gamma \not\geq 0$. We note that

$$\big[\Gamma M(\theta)\big]_{i,j} = \sum_{k,j',j''=1}^{N} \gamma(i,k)P_1^{-1}(k,j')P_2(j'',j)Q(j',j'')e^{T\left(\theta\lambda_1^{(j')}+(1-\theta)\lambda_2^{(j'')}\right)}.$$

**Log-convexity of the spectral radius.** A celebrated result of Kingman [134] asserts that if the entries of a nonnegative matrix are log convex functions of a variable then so is the spectral radius of the matrix. If this property applies to the positive matrix $M(\theta)$, $\theta \mapsto \rho(M(\theta))$ is log-convex. In this case, it is monotone (yielding (SSTP)) provided that the derivatives at 0 and 1 have the same sign, that is:

$$\big(\mu_2 - \langle DF^1(0)V^2, V_*^2\rangle\big)\big(\langle DF^2(0)V^1, V_*^1\rangle - \mu_1\big) > 0, \tag{11.18}$$

where $\mu_i = \mu(DF^i(0))$, and $V^i$ (resp. $V_*^i$) is the principal eigenvector of $DF^i(0)$ (resp. of $DF^i(0)^*$) with $V^i, V_*^i \gg 0$ and $\langle V^i, V^i\rangle = 1 = \langle V^i, V_*^i\rangle$.

When $DF^i(0)$ are diagonalizable ($i \in \{1, 2\}$), the above formula shows that

$$M_{i,j}(\theta) = \sum_{n=1}^{N^2} \alpha_n(i,j)e^{\beta_n(i,j)\theta}$$

for some $\alpha, \beta$. In cases when $M_{i,j}$ can be proved to be a log-convex function of $\theta$, (SSTP) holds under assumption (11.18).

**Computation of the second-order derivative.** A more general condition for (SSTP) than the monotonicity of $\rho$ would be that $\rho$ is either concave or convex (or log-concave, or log-convex). To formulate this condition we compute the second-order derivative of $\log(\rho)$ from (11.12) as

$$\frac{d}{d\theta}\big(\log(\rho(\theta))\big) = r'(\theta) = \underbrace{\langle SV'(\theta), V_*(\theta)\rangle}_{=:R_1} + \underbrace{\langle SV(\theta), V_*'(\theta)\rangle}_{=:R_2},$$

where

$$S = DF^1(0) - DF^2(0). \tag{11.19}$$

Differentiating with respect to $\theta$ the eigenvector equations for $V(\theta)$ and $V_*(\theta)$ along with their normalizations $\langle V(\theta), V(\theta)\rangle = 1$ and $\langle V(\theta), V_*(\theta)\rangle = 1$ yields:

$$\begin{aligned}(M(\theta) - \rho(\theta)I)V'(\theta) &= (\rho'(\theta)I - M'(\theta))V(\theta),\\ (M^*(\theta) - \rho(\theta)I)V_*'(\theta) &= (\rho'(\theta)I - (M^*)'(\theta))V_*(\theta),\\ \langle V(\theta), V'(\theta)\rangle = 0 &= \langle V'(\theta), V_*(\theta)\rangle + \langle V(\theta), V_*'(\theta)\rangle.\end{aligned}$$

Dropping the argument $\theta$, we note that $V', V_*'$ are well-defined from these linear equations since $\mathrm{Im}(M - \rho I) = (V_*\mathbb{R})^\perp$ (and symmetrically $\mathrm{Im}(M^* - \rho I) = (V\mathbb{R})^\perp$) and the scalar product conditions give uniqueness. We introduce the notation $H := (V\mathbb{R})^\perp$ (resp. $H_* := (V_*\mathbb{R})^\perp$) for the

hyperplane with normal vector $V$ (resp. $V_*$). We also introduce the Perron projection operator $\Pi := V_* V^* \in \mathcal{L}(\mathbb{R}^N)$, and its adjoint $\Pi^* = V V_*{}^*$.

In particular, $M - \rho I \in \mathcal{L}(H, H_*)$ is an invertible linear application, whose inverse is denoted $M_r \in \mathcal{L}(H_*, H)$, and we have

$$V' = M_r\big((\rho' I - M')V\big).$$

Symmetrically, $M^* - \rho I \in \mathcal{L}(H)$ is invertible (since $V_* \notin H$), its inverse is denoted $M_a^* \in \mathcal{L}(H)$ and

$$V_*' = M_a^*\big((\rho' I - M_*')V_*\big) - \langle V_*, V'\rangle V_*.$$

Using the notation $M_i = DF^i(0)$ ($i \in \{1,2\}$), from the definition $M(\theta) = e^{T(1-\theta)M_2} e^{T\theta M_1}$ we also have:

$$M' = T(MM_1 - M_2 M), \tag{11.20}$$

$$(M^*)' = T\big((M_1)^* M^* - M^*(M_2)^*\big). \tag{11.21}$$

In order to compute the two terms in $r'$, we note two preliminary identities. First, using (11.20) and (11.12) we get

$$\frac{1}{T}(\rho' I - M')V = \rho(\Pi^* - I)SV - (M - \rho I)M_1 V, \tag{11.22}$$

where both terms in the right-hand side belong to $H_*$. Symmetrically, using (11.21) and (11.12) we get

$$\frac{1}{T}(\rho' I - (M^*)')V_* = (M^* - \rho I)M_2^* V_* + \rho(\Pi - I)S^* V_*, \tag{11.23}$$

where both terms in the right-hand side belong to $H$.

Then, using (11.22), $M_r \in \mathcal{L}(H_*, H)$ and $M_r \circ (M - \rho I) = I_H$ we can compute

$$\begin{aligned} R_1 &= \big\langle M_r\big((\rho' I - M')V\big), S^* V_*\big\rangle, \\ &= T\rho\big\langle M_r(\Pi^* - I)SV, S^* V_*\big\rangle - T\langle M_1 V, S^* V_*\rangle. \end{aligned}$$

Symmetrically, using (11.23), $M_a^* \in \mathcal{L}(H)$ and $M_a^* \circ (M^* - \rho I) = I_H$ we obtain

$$\begin{aligned} R_2 &= \langle SV, M_a^*\big((\rho' I - M_*')V_*\big) - \langle V_*, V'\rangle V_*\rangle, \\ &= T\rho\big\langle SV, M_a^*(\Pi - I)S^* V_*\big\rangle + T\langle SV, M_2^* V_*\rangle - \langle SV, V_*\rangle\langle V_*, V'\rangle. \end{aligned}$$

Using (11.22) with $M_r \in \mathcal{L}(H_*, H)$ and $(M - \rho I) \circ M_r = I_H$ we also get

$$\begin{aligned} \langle V_*, V'\rangle &= \langle V_*, M_r\big((\rho' I - M')V\big)\rangle, \\ &= T\rho\big\langle V_*, M_r(\Pi^* - I)SV\big\rangle - T\langle V_*, M_1 V\rangle. \end{aligned}$$

Gathering $R_1$ and $R_2$ we obtain

$$\frac{r'}{T} = \overbrace{\big(\langle SV, V_*\rangle\big)^2 + \big\langle (M_2 S - S M_1)V, V_*\big\rangle}^{r_1} + \underbrace{\rho\big\langle M_r(\Pi^* - I)SV, (S^* - \langle SV, V_*\rangle I)V_*\big\rangle}_{r_2} + \underbrace{\rho\big\langle M_a^*(\Pi - I)S^* V_*, SV\big\rangle}_{r_3}.$$

We notice that

$$r_2 = \rho\big\langle M_r(\Pi^* - I)SV, (I - \Pi)S^* V_*\big\rangle = \rho\big\langle SV, (I - \Pi)M_r^*(\Pi - I)S^* V_*\big\rangle$$

and

$$r_3 = \rho\big\langle SV, M_a^*(\Pi - I)S^* V_*\big\rangle,$$

so $r_2 = r_3$, since $(M^* - \rho I) \circ M_a^* = I_H$, $(M^* - \rho I) \circ M_r^* = I_H$ and $(M^* - \rho I) \circ \Pi M_r^* = 0$

Finally $\rho'' = T^2 \rho r^2 + T\rho r'$ whence

$$\frac{\rho''}{T^2 \rho} = 2\big(\langle SV, V_*\rangle\big)^2 + \big\langle (M_2 S - S M_1)V, V_*\big\rangle + 2\rho\big\langle M_a^*(\Pi - I)S^* V_*, SV\big\rangle. \tag{11.24}$$

In principle, the identity (11.24) could be used to derive (SSTP) under more general conditions on $M_1 = DF^1(0), M_2 = DF^2(0)$ than those given in Theorem 11.2. However, we do not explore such conditions in the present article.

**Time scaling.** Until now we have considered that the period $T > 0$ was fixed. Letting $T$ go to 0 or $+\infty$ yields interesting limits. For an irreducible Metzler matrix $U$,

$$e^{-T\mu(U)}e^{TU} \xrightarrow[T\to+\infty]{} VV_*^*$$

where $V$ is the principal eigenvector of $U$ and $V_*$ is the principal eigenvector of $U^*$, normalized by $V_*^*V = 1$. From this fact, we have

$$e^{-T(\theta\mu(DF^1(0))+(1-\theta)\mu(DF^2(0))}M(\theta) \xrightarrow[T\to+\infty]{} V(0)V_*(0)^*V(1)V_*(1)^*,$$

from which we deduce that

$$\frac{1}{T}\log(\rho(\theta)) \sim_{T\to+\infty} \theta\mu(DF^1(0)) + (1-\theta)\mu(DF^2(0)).$$

In fact, we even get the next term in the asymptotic development:

$$\log(\rho(\theta)) - T\big(\theta\mu(DF^1(0)) + (1-\theta)\mu(DF^2(0))\big) - \log\big(V_*(0)^*V(1)V_*(1)^*V(0)\big) = o_{T\to\infty}(1).$$

Therefore, for $T$ large enough, $\rho$ is close to be monotone, and even close to be equal to the exponential interpolation of $T\mu(DF^1(0))$ and $T\mu(DF^2(0))$.

Meanwhile, $\lim_{T\to 0}\rho(\theta) \equiv 1$.

**Optimization problems.** For a general two-seasonal model defined by a monotone and concave map $G : \mathcal{P}\times\mathbb{R}^N \to \mathbb{R}^N$ and $\pi^U, \pi^F \in \mathcal{P}$, a natural question is the optimization of the spectral radius when the favorable and unfavorable seasons can be split throughout the year. Let $M_\sharp := T\cdot DG(\pi_\sharp, 0)$ (with $\sharp \in \{U, F\}$). For $K \in \mathbb{Z}_+$, we define:

$$\overline{\rho}_{M_U,M_F}(\theta, K) = \max_{(\sigma,\sigma')\in\varphi_K(\theta)} \rho(M_{M_U,M_F}(\sigma,\sigma')), \tag{11.25}$$

$$\underline{\rho}_{M_U,M_F}(\theta, K) = \min_{(\sigma,\sigma')\in\varphi_K(\theta)} \rho(M_{M_U,M_F}(\sigma,\sigma')), \tag{11.26}$$

where

$$\varphi_K(\theta) := \Big\{\big((\theta_k)_k, (\theta'_k)_k\big) \in [0,1]^{2K}, \sum_{k=1}^K \theta_k = \theta, \sum_{k=1}^K \theta'_k = 1-\theta\Big\}$$

is compact and for $(\sigma,\sigma') \in \varphi_K(\theta)$ and $M_1, M_2 \in M_N(\mathbb{R})$,

$$M_{M_1,M_2}(\sigma,\sigma') := e^{\theta'_K M_2}e^{\theta_K M_1}\cdots e^{\theta'_1 M_2}e^{\theta_1 M_1}.$$

Note that by Gelfand's formula,

$$\rho(M(\sigma,\sigma')) \le \prod_k \rho(e^{\theta'_k M_2})\rho(e^{\theta_k M_1}) = e^{\theta\mu_1+(1-\theta)\mu_2},$$

where $\mu_i = \mu(M_i)$.

**Remark 11.6.** *In the specific case when $M_U$ and $M_F$ are irreducible Metzler matrices with the same principal eigenvector (that is, condition (A)) , $\rho(M(\sigma,\sigma'))$ does not depend on $(\sigma,\sigma') \in S_K(\theta)$ and does even not depend on $K \in \mathbb{Z}_+$: we have*

$$\forall K \in \mathbb{Z}_+, \forall\theta\in[0,1], \quad \overline{\rho}_{M_U,M_F}(\theta,K) = e^{\left(\theta\mu_U+(1-\theta)\mu_F\right)} = \underline{\rho}_{M_U,M_F}(\theta,K),$$

*with $\mu_\sharp = \mu(M_\sharp)$.*

*In this case, assuming $\mu_F > 0 > \mu_U$ we recover Theorem 11.3 with*

$$\theta_* = \frac{\mu_F}{\mu_F - \mu_U}.$$

**Acknowledgements.** The authors wish to thank Dongmei Xiao and Jean-Pierre Françoise for useful discussions, and Benoit Perthame for valuable comments which helped to improve this manuscript. Part of this work was done while HJ was visiting Dongmei Xiao at SJTU.

# Appendices

## 11.A Proof of Theorem 11.1

We consider the following $T$-periodic piecewise-autonomous differential equation

$$\frac{dx}{dt} = F(t,x), \tag{11.27}$$

where for all $x \in \mathbb{R}^N$, $F(\,\cdot\,,x)$ is a piecewise-constant function. We assume that there is a family of functions $(F^k)_k : \mathbb{R}^N_+ \to \mathbb{R}^N_+$ such that:

$$F(t,x) = F^k(x) \text{ if } \frac{t}{T} - \Big\lfloor \frac{t}{T} \Big\rfloor \in [\theta_{k-1}, \theta_k)$$

where $(\theta_i)_{0\le i\le N} \in [0,1]^{N+1}$ is a non-decreasing family such that $\theta_0 = 0$ and $\theta_N = 1$. For $x \in \mathbb{R}$, the notation $\lfloor x \rfloor$ stands for the largest integer $n \in \mathbb{Z}$ such that $n \le x$.

We assume that for all $1 \le k \le K$, $F^k : \mathbb{R}^N_+ \to \mathbb{R}^N_+$ is continuously differentiable, monotone (that is, if $x \ll y$ then $F^k(x) \ll F^k(y)$), concave (that is, if $x \ll y$ then $DF^k(x) \gg DF^k(y)$) and satisfies $F^k(0) = 0$.

Following the lines of [207] and [126], to prove Theorem 11.1 we split into four assertions the various hypotheses of [207, Theorem 2.1], to check that they hold for the Poincaré map for (11.27). We begin with:

**Lemma 11.5.** *If $x(t)$ is a solution of (11.27) with $x(t_0) \ge 0$, then $x(t)$ can be extended to $[t_0, +\infty]$ and $x(t) \ge 0$ for $t \ge t_0$.*

*Proof.* Let $t \ge 0$. For all $y \ge 0$, by concavity of all $F^k$ $(1 \le k \le K)$, we have $D_x F(t,y) \le D_x F(t,0)$. Hence for all $t \ge 0$ and $x \ge 0$,

$$\begin{aligned} F(t,x) &= F(t,0) + \Big( \int_0^1 D_x F(t,sx)ds \Big) x \\ &\le F(t,0) + D_x F(t,0)x \text{ since } x \ge 0. \end{aligned}$$

Let $y$ be the solution to the affine differential equation $y' = F(t,0) + D_x F(t,0)y$, $y(t_0) = x(t_0)$. From Kamke's theorem, we deduce that $x(t) \le y(t)$ on the maximal interval of existence $[t_0, w)$ of $x(t)$. Since $y(t)$ is defined for all $t \ge t_0$, it follows that $w = +\infty$.

The standard positivity property (P) implies $x(t) \ge 0$ for $t \ge t_0$. □

Then, as an immediate consequence of monotonicity and Kamke's theorem:

**Lemma 11.6.** *If $x(t)$ and $y(t)$ are solutions of (11.27) with $0 \le y(t_0) \ll x(t_0)$, then $y(t) \ll x(t)$ for $t > t_0$.*

For all $s \in \mathbb{R}$ and $x_0 \in \mathbb{R}^N$, we denote by $t \mapsto \phi(t;s,x_0)$ the solution of (11.27) which satisfies $x(s) = x_0$. In particular, $\phi(s;s,x) = x$. For all $1 \le k \le K$, we also introduce $t \mapsto \phi^k(t;s,x_0)$ as the solution to

$$\frac{dx}{dt} = F^k(x), \quad x(s) = x_0.$$

By regularity of $F^k$, each $\phi^k(\theta_k T, \theta_{k-1}T, \,\cdot\,)$ is a $C^1$ function.

With these notations it follows from Lemmas 11.5 and 11.6 that the Poincaré map

$$P(x) := \phi(T;0,x) = \phi^K\big(\theta_K T; \theta_{K-1}T, \phi^{K-1}\big(\cdots \phi^1(\theta_1 T;0,x)\big)\big), \quad x \ge 0 \tag{11.28}$$

is well defined as a $C^1$ map $P : \mathbb{R}_+^N \to \mathbb{R}_+^N$ because it is a composition of functions of class $C^1$. In order to apply [207, Theorem 2.1], we must verify that the differential $DP$ satisfies:

$$DP(0) \gg 0 \text{ and } DP(x) \geq 0 \text{ if } x \gg 0, \tag{$M_0$}$$
$$DP(y) < DP(x) \text{ if } 0 \ll x \ll y. \tag{$C_0$}$$

Introducing the notations, for $x \in \mathbb{R}^N$

$$\widetilde{\phi}^k(x) := \phi^k\big(\theta_k T; \theta_{k-1}T, \widetilde{\phi}^{k-1}(x)\big) \in \mathbb{R}^N \text{ for } 1 \leq k \leq K, \quad \widetilde{\phi}_0(x) := x,$$
$$\widehat{\phi}^k(x) := \frac{\partial \phi^k}{\partial x}(\theta_k T; \theta_{k-1}T, x) \in \mathbb{R}^{N\times N},$$

we can compute

$$DP(x) = \frac{\partial \phi}{\partial x}(T; 0, x) = \prod_{k=1}^{K} \widehat{\phi}^k \circ \widetilde{\phi}^{k-1}(x). \tag{11.29}$$

We write $\Phi(t, x) := \frac{\partial \phi}{\partial x}(t; 0, x)$, so that $DP = \Phi(T, \cdot)$. By construction, $\Phi(t, x)$ is the fundamental matrix for the variational equation

$$X' = D_x F(t, \phi(t; 0, x))X, \quad X(0) = I \tag{11.30}$$

where $I$ is the $N \times N$ identity matrix. Lemma 11.7 below is a direct consequence of (M)

**Lemma 11.7.** *If $x \gg 0$, then $\Phi(t, x) > 0$ for $t > 0$. In addition, $\Phi(t, 0) \gg 0$ for $t > 0$.*

*Proof.* Let $T > 0$ and $x \in \mathbb{R}^N$. Let $M = M_{T,x} \in (0, +\infty)$ such that $D_x F(t, \phi(t; 0, x)) + MI \geq 0$ for all $t \in [0, T]$. As long as $\Phi(t, x) \geq 0$ on $[0, T]$ we have on this interval $\frac{d}{dt}\Phi(t, x) \geq -M\Phi(t, x)$, hence $\Phi(t, x) \geq e^{-Mt} I > 0$.

Then, $\Phi(t, 0)$ solves (11.3) with $\Phi(0, 0) = I$. Since $D_x F(t, 0)$ is an irreducible (by (I)) Metzler matrix, $\Phi(t, 0) \gg 0$ for $t > 0$. □

Applying Lemma 11.7 with $t = T$ yields ($M_0$). It remains only to verify ($C_0$), which is the object of the next lemma

**Lemma 11.8.** *If $0 \ll x \ll y$, then $DP(x) > DP(y)$.*

*Proof.* We write $Z(t, x) = D_x F(t, \phi(t; 0, x))$ for short. If $0 \ll x \ll y$, from Lemma 11.6, we have $\phi(t; 0, x) \ll \phi(t; 0, y)$ for all $t \geq 0$. By (C), we deduce that $Z(t, x) > Z(t, y)$. Hence

$$\begin{aligned}\Phi'(t, x) &= Z(t, x)\Phi(t, x) \\ &\geq Z(t, y)\Phi(t, x),\end{aligned}$$

since $\Phi(t, x) \geq 0$ by Lemma 11.7. Therefore, it follows from Kamke's theorem that $\Phi(t, x) \geq \Phi(t, y)$.

Then, we follow [14, Lemma 1] by letting $Y(t) = \Phi(t, x) - \Phi(t, y)$. $Y(t)$ satisfies

$$Y'(t) = Z(t, x)Y(t) + [Z(t, x) - Z(t, y)]\Phi(t, y), \quad Y(0) = 0.$$

Using the fundamental matrix $\Phi$ we get

$$Y(T) = \int_0^T \Phi(T, x)\Phi(s, x)^{-1}[Z(s, x) - Z(s, y)]\Phi(s, y)ds$$

Now, $Z(t, s) \equiv \Phi(t, x)\Phi(s, x)^{-1} > 0$ for $t > s$ since it is the fundamental matrix at $t = s$ of $z' = Z(t, x)z$ (exactly as in Lemma 11.7). Since $\Phi(s, y) > 0$ for $0 < s \leq T$ and $Z(s, x) - Z(s, y) \gg 0$ for $0 \leq s \leq T$, it follows that $Y(T) > 0$. This is the desired conclusion. □

We have verified all assumptions and can apply [207, Theorem 2.1] and Theorem 11.1 follows immediately on noting that $\lambda = \rho(DP(0)) = \rho(\Phi(T, 0))$ is the characteristic multiplier of (11.3) of maximum modulus.

# Part IV

# Other aspects

# Chapter 12

# Selection-mutation dynamics with sexual reproduction

> Alors même que tous les hommes apprennent à parler et la plupart à lire, les enfants continuent à naître en ne sachant ni parler ni lire.
>
> Michel Tournier, *Le Miroir des idées.*

This chapter is a joint work with Cécile Taing and Benoît Perthame. An earlier version has appeared in the PhD thesis of Cécile Taing [214, Chapter 3].

**Abstract.** We study a family of selection-mutation models of a sexual population structured by a phenotypical trait. The main feature of these models is the asymmetric trait heredity or fecundity between the parents: we assume that each individual inherits mostly its trait from the female or that the trait modifies the female fecundity but not the male one. Following previous works inspired from principles of adaptive dynamics, we rescale time and assume that mutations have limited effects on the phenotype. Our goal is to study the asymptotic behavior of the population distribution. We derive non-extinction conditions and BV estimates on the total population. We also obtain Lipschitz estimates on the solutions of Hamilton-Jacobi equations that arise from the study of the population distribution concentration at fittest traits. Concentration results are obtained in some special cases by using a Lyapunov functional.

## 12.1 Introduction

We introduce and study mathematically a family of models of selection-mutation for a continuous phenotype, which we call "trait", denoted by $x \in \mathcal{P}$. The set of phenotypes $\mathcal{P}$ is a complete metric space, typically $\mathcal{P} = \mathbb{R}$. We assume that all individuals compete for survival because they share the same resources. This assumption implies the boundedness of the total population.

Although our approach is formal and mathematical, the models under study are motivated by the issue of insecticide resistance. This phenomenon has appeared in many insects of interest for human health, in particular in species of mosquitoes that are vectors for dengue (in the *Aedes* genus) or malaria (in the *Anopheles* genus). For this specific problem of selection-mutation, the trait variable should contain, for instance, the expression level for the *kdr* gene (*knock-down resistance*, see [185]). The present study is part of a more general program on the analysis of models, and their control, in the context of evolutionary epidemiology (see [36, 176, 211, 212]).

Because of this motivation, and as a new feature, our models have a sexual reproduction kernel. This is not the case in similar selection-mutation models developed for bacteria or resistance to treatment in cancer (see [153]), where the reproduction is clonal. The major feature of equations for sexual reproduction is to yield nonlinear and nonlocal birth terms with a quadratic aspect though 1-homogeneous. All models studied in the present paper are derived from the general form

$$\begin{cases} \epsilon \partial_t n_\epsilon(t,x) = \frac{1}{\rho_\epsilon(t)} \iint K_\epsilon(x,y,z) n_\epsilon(t,y) n_\epsilon(t,z) dy\, dz - R(x,\rho_\epsilon(t)) n_\epsilon(t,x), \\ \rho_\epsilon(t) = \int_{\mathcal{P}} n_\epsilon(t,x) dx, \quad n_\epsilon(0,x) = n_\epsilon^0(x). \end{cases} \tag{12.1}$$

The variable $t$ stands for time, $n_\epsilon(t,x) \in [0,+\infty)$ is the population density at time $t$ and trait $x$ and $\rho_\epsilon(t)$ is the total population. The positive function $R$ is here the saturation term, which contains the death rate and the insecticide effect. Competition is taken into account through the dependency of $R$ in its second variable.

In equation (12.1), we interpret $y$ (the second variable for $K_\epsilon$) as the female trait, and $z$ (the third variable) as the male trait. Thus $x \mapsto K_\epsilon(x,y,z)$ is equal to the distribution of individuals that are born from any encounter between a female of trait $y$ and a male of trait $z$, per unit of time. Of course, this model is valid only if we assume that the sex ratio is constant in time and the same for each value of the trait. We make this simplification in order to obtain a single equation rather than a system.

It is worth highlighting that we aim here at general properties and methods for dealing with the nonlinear and nonlocal birth term, rather than at a realistic model for the evolution of a specific trait (see [219, 221] for more realistic models). We hope that the techniques developed here will be successfully applied to specific contexts. We also point out that the same kind of equation structure appears in models of cell population exchanging genetic information (see [33, 159]) or proteins (see [160]).

The relationships between sexual selection and speciation are not well understood. Models of sexual reproduction have already been discussed in different contexts. Studies of individual-based models of sexual population were performed to determine the necessary conditions to evolutionary branching in [65, 136, 226]. Mendelian populations, *i.e.* structured by genetic types, were also considered (see [41, 42] and the discussion in Section 13.4 below). In [57] for instance, the authors investigate a stochastic birth and death process model for sexually reproducing diploids with Lotka-Volterra type dynamics and single locus genetics. At the small mutation steps limit, they derive a differential equation in allele space, referred to as a form of the canonical equation of the adaptive dynamics. In [59], another stochastic birth and death process model is studied with sexual reproduction according to mating preferences and a space structure with patches. In this case, reproductive isolation between patches occurs, and the authors prove that the time needed for this isolation to occur is a function of the population size. In [200], a deterministic system with three phenotypes (two alleles at a single locus) was studied for which the "reversal time" (measuring the persistence of resistance in a population after exposition to insecticide) was studied.

From a full population point of view, in [171] the authors considered sexual populations structured by a trait and a space variable in a non-homogeneous environment, and after performing an asymptotic limit and a simplification of the model, they derived an estimate of the invasion speed or extinction speed of the population. In [37], the authors study the same kind of models as in the present paper, where the traits of the newborns are distributed through a gaussian kernel centered on the mean of the parents' traits and with a constant variance, as in [68]. They prove the existence of principal eigenelements for the corresponding eigenproblem, using the Schauder fixed point theorem.

The main results of this paper regard the behavior of $\rho_\epsilon$ and $n_\epsilon$ in the asymptotic of large time scale and mutations with limited effect on the phenotype, for several models of the form (12.1). We also identify some difficulties raised by the application of our methods to the general case of (12.1).

In the present paper, we study two classes of models with the common idea that new individuals inherit mostly their trait from the female. We consider a first model with **asymmetric fecundity** (AF in short)

$$\epsilon\partial_t n_\epsilon(t,x) = \frac{1}{\rho_\epsilon(t)} \iint B(y)\alpha_\epsilon(x,y,z)n_\epsilon(t,y)n_\epsilon(t,z)dy\,dz - R(x,\rho_\epsilon(t))n_\epsilon(t,x), \qquad \text{(AF)}$$

where $B$ is a positive function (crossing fecundity, which is assumed to depend only on the female's trait), $\alpha_\epsilon(\cdot\,,y,z)$ is the probability distribution of the offspring from a $y$ female and a $z$ male,

$$K_\epsilon(x,y,z) = B(y)\alpha_\epsilon(x,y,z), \quad \int \alpha_\epsilon(x,y,z)\,dx = 1 \text{ for all } y,z.$$

The second model features an **asymmetric trait heredity** (ATH in short) with $\mathcal{P} = \mathbb{R}$, which reads

$$\epsilon\partial_t n_\epsilon(t,x) = \frac{1}{\rho_\epsilon(t)} \iint K_0(x-z)G_\epsilon(x-y)n_\epsilon(t,y)n_\epsilon(t,z)dy\,dz - R(x,\rho_\epsilon(t))n_\epsilon(t,x), \qquad \text{(ATH)}$$

where

$$K_\epsilon(x,y,z) = K_0(x-z)G_\epsilon(x-y), \qquad (12.2)$$

with $K_0, G_\epsilon$ positive functions such that there exists a positive function $G$, $\int G(z)dz = 1$ and $G_\epsilon(x-z) = \frac{1}{\epsilon}G\left(\frac{x-z}{\epsilon}\right)$.

We study two ingredients of proof for convergence: first, we identify a consistent limit object as $\epsilon \to 0$, which is here a constrained Hamilton-Jacobi equation; secondly we obtain time compactness estimates on the solutions at the $\epsilon$-level in order to be able to extract converging subsequences and to use the stability property of viscosity solutions. The first and most intricate step to obtain this second ingredient is the study of the variations of $\rho_\epsilon$.

For simplification, we first study a model without mutations, which is a particular case of the two models presented above, in order to introduce the ingredients that we use and to highlight the new arguments of the proofs. This model with no mutations, posed on $\mathcal{P} = \mathbb{R}$ reads

$$\epsilon\partial_t n_\epsilon(t,x) = \left(\frac{1}{\rho_\epsilon(t)}K_0 * n_\epsilon(t,\cdot)(x) - \nu\rho_\epsilon(t)\right) n_\epsilon(t,x), \tag{nM}$$

with $\nu > 0$. This equation can be written under the form of equation (AF) with

$$B \equiv 1, \quad R(x,\rho) \equiv \nu\rho \text{ and } \alpha_\epsilon(x,y,z) = K_0(x-z)\delta_0(x-y),$$

and also under the form of (ATH) with

$$G_\epsilon = \delta_0.$$

We also generalize (nM) to any (complete metric) phenotype space $\mathcal{P}$ with

$$\epsilon\partial_t n_\epsilon(t,x) = \Big(\frac{1}{\rho_\epsilon(t)}\int K(x,y)n_\epsilon(t,y)dy - \big(R_0(x) + R_1(\rho_\epsilon)\big)\Big)n_\epsilon(t,x), \quad n_\epsilon(0,x) = n^0(x), \tag{gnM}$$

for some symmetric kernel $K : \mathcal{P}^2 \to \mathbb{R}_+$, and obtain Lyapunov convergence results for (gnM).

The paper is organized as follows. In Section 12.2, we state our assumptions and results. We also establish some non-extinction conditions and bounds on the total population. In Section 12.3, we focus on the models without mutations (nM)-(gnM) in order to introduce the main arguments that will be used for the more general cases. In particular we derive $BV$ estimates for the total population and discuss the formal limit of the population distribution. In Section 12.4, we address the derivation of $BV$ estimates for the (ATH) and (AF) models when $R$ only depends on the total population variable and we explain the difficulties encountered when $R$ is generic. In Section 12.5 we briefly explain the settings leading to Lyapunov convergence results. In Section 12.6, we deal with the Hamilton-Jacobi approach.

## 12.2 Main results

### 12.2.1 Assumptions and statements

The function $R$ stands for the death rate and the competition effects. We make the standing assumption that it increases with the total population:

$$\forall x, \rho, \quad \partial_\rho R(x,\rho) > 0. \tag{12.3}$$

Some important results are obtained when $R$ has the very simple form

$$R(x,\rho) = \nu\rho, \quad \forall x \in \mathbb{R}, \text{ with } \nu > 0. \tag{12.4}$$

We also assume usually that the initial data satisfies

$$\epsilon(\dot\rho_\epsilon)_-(0) = \left(\int n^0_\epsilon(x)\frac{K_0 * n^0_\epsilon}{\rho^0_\epsilon}(x)\,dx - (\rho^0_\epsilon)^2\right)_- \text{ is uniformly bounded,} \tag{12.5}$$

where for $a \in \mathbb{R}$ we use the notation $a_- = \max(-a, 0)$.

For models with no mutations (nM) and asymmetric trait heredity (ATH), we assume

$$K_0 \in \mathcal{C}_b(\mathbb{R}, \mathbb{R}_+) \text{ is a symmetric kernel,} \tag{12.6}$$

where $\mathcal{C}_b(\mathbb{R}, \mathbb{R}_+)$ is the space of continuous and bounded functions $\mathbb{R} \to \mathbb{R}_+$.

We state the following result for (nM), whose proof is given in Section 12.3.

**Theorem 12.1** (*BV* bound for model (nM)). *Let $T > 0$ and let $n_\epsilon$ be the solution to* (nM) *associated with initial data $n_\epsilon^0$. We assume* (12.5) *and* (12.6).

*Then, $\rho_\epsilon$ is uniformly bounded in $BV(0,T)$. Namely, we obtain*

$$\int_0^T |\dot\rho_\epsilon(t)|\,dt \le \rho_M + 2(\dot\rho_\epsilon)_-(0)\frac{\epsilon}{\kappa_m''}\big(1-e^{-\frac{\kappa_m'' T}{\epsilon}}\big),$$

*with $\rho_M$ and $\kappa_m''$ defined later on. This implies that, up to extraction of subsequences, there exist limits $\rho_\epsilon \to \rho$ in $L^1(0,T)$, and $n_\epsilon \rightharpoonup n \in L^\infty_t(0,T;\mathcal{M}_+(\mathbb{R}))$ in the sense of measures.*

*Moreover, we have*

$$\int_0^T \int_{\mathbb{R}} n_\epsilon \Big(\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\Big)^2 dx\,dt = O(\epsilon). \tag{12.7}$$

For the model with asymmetric fecundity (AF), we need the following assumption on $B$ and $\alpha$:

$$\begin{aligned}&\exists C>0, \forall \epsilon>0, \forall \phi \in \mathcal{M}_+^1(\mathcal{P}),\\ &\iiint \alpha_\epsilon(x,y,z)B(x)B(y)\phi(y)\phi(z)dx\,dy\,dz - \Big(\int B(y)\phi(y)dy\Big)^2 \ge -C\epsilon.\end{aligned} \tag{12.8}$$

This means that the fecundity variation from one generation to the next is controlled and in fact is non-decreasing as $\epsilon$ goes to 0. We obtain the following result whose proof is given in Section 12.4.

**Theorem 12.2** (*BV* bound for (AF)). *Let $T > 0$ and let $n_\epsilon$ be the solution to* (AF) *associated with initial data $n_\epsilon^0$. Assume* (12.4) *and* (12.8).

*Then, $\rho_\epsilon$ is uniformly bounded in $BV(0,T)$. Namely, we have*

$$\int_0^T |\dot\rho_\epsilon(t)|\,dt \le \rho_M + 2(\dot\rho_\epsilon)_-(0)\frac{\epsilon}{\nu\rho_m}\big(1-e^{-\frac{\nu\rho_m T}{\epsilon}}\big) + 2\frac{C}{\nu\rho_m}\left(T + \frac{\epsilon}{\nu\rho_m}(e^{-\frac{\nu\rho_m T}{\epsilon}}-1)\right).$$

*with $C$, $\rho_M$ and $\rho_m$ defined later on. This implies that, up to extraction of subsequences, there exist limits $\rho_\epsilon \to \rho$ in $L^1(0,T)$, and $n_\epsilon \rightharpoonup n \in L^\infty_t(0,T;\mathcal{M}_+(\mathcal{P}))$ in the sense of measures.*

For (ATH), we also need to assume that, for all $\phi \in L^1 \cap W^{1,\infty}$, $G_\epsilon * \phi = \phi + O(\epsilon)$, in the sense

$$\frac{1}{\epsilon\|\phi\|_{Lip}}\|G_\epsilon * \phi - \phi\|_{L^1} \text{ is uniformly bounded in } \epsilon. \tag{12.9}$$

Additionally, we assume that $K_0$ is Lipschitz in this case. We obtain the following result whose proof is given in Section 12.4.

**Theorem 12.3** (*BV* bound for (ATH)). *Let $n_\epsilon$ be the solution to* (ATH) *associated with initial data $n_\epsilon^0$. Assume* (12.5), (12.6) *and* (12.9). *Assume also the following ("non-extinction" in this case) condition*

$$\exists \eta_0 > 0, \quad \forall \epsilon > 0, \quad \eta_\epsilon := \inf_{\phi \in \mathcal{M}_+^1(\mathbb{R})} \int K_0 * \phi \cdot G_\epsilon * \phi\,dx \ge \eta_0. \tag{12.10}$$

*Then $\rho_\epsilon$ is uniformly bounded in $BV(0,T)$. Namely, we have*

$$\int_0^T |\dot\rho_\epsilon(t)|\,dt \le \rho_M + 2(\dot\rho_\epsilon(0))_-\frac{\epsilon}{C_1}(1-e^{-C_1T/\epsilon}) + 2\frac{\epsilon C_2}{C_1^2}(e^{-C_1T/\epsilon}-1) + 2\frac{C_2}{C_1}T.$$

*Then, up to extraction there exist $\rho \in L^1_{loc}(0,\infty)$ and $n \in L^\infty_t(0,T;\mathcal{M}_+(\mathbb{R}))$ such that $(\rho_\epsilon)$ converges towards $\rho$ in $L^1_{loc}(0,\infty)$, and $(n_\epsilon)$ towards $n$ in the sense of measures, when $\epsilon$ goes to 0.*

*Moreover, for all $T > 0$, we have*

$$\int_0^T \int_{\mathbb{R}} (G_\epsilon * n_\epsilon)\left[\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\right]^2 dx\,dt = O(\epsilon),$$

In the general case of a death rate depending on both traits and the total population, we can perform the Hopf-Cole transform

$$u_\epsilon(t,x) = \epsilon \ln n_\epsilon(t,x),$$

and apply a Hamilton-Jacobi approach. For the models under investigation, we obtain the following result that is the topic of Section 12.6.

**Theorem 12.4** (Lipschitz estimates for $u_\epsilon$). *Under some assumptions on the initial data $u_\epsilon^0$, for both models (AF) and (ATH), the corresponding $u_\epsilon$ are locally Lipschitz uniformly in $\epsilon$.*

*Moreover, we have a global upper bound on $u_\epsilon$. Namely, there exists a constant $C$, such that*

$$u_\epsilon(t,x) \le \epsilon \ln\left(C + \frac{C(1+t)}{\epsilon}\right).$$

The assumptions required for the proof of this theorem in each case are specified in the corresponding section.

### 12.2.2 Boundedness of $\rho_\epsilon$ and non-extinction

The total population $\rho_\epsilon$ satisfies

$$\epsilon\dot\rho_\epsilon(t) = \int \Big( \iint K_\epsilon(x,y,z)\frac{n_\epsilon(t,z)}{\rho_\epsilon(t)} n_\epsilon(t,y)dy\,dz - R(x,\rho_\epsilon(t))n_\epsilon(t,x)\Big)\,dx.$$

To ensure that $\rho_\epsilon$ remains bounded along all trajectories, we complement (12.3) with

$$\begin{aligned} &\exists R_m : \mathbb{R}_+ \to \mathbb{R}_+, \text{ increasing, with } R_m(0) = 0,\\ &R_m(+\infty) = +\infty \text{ and } \forall x, \quad R(x,\rho) \ge R_m(\rho). \end{aligned} \tag{12.11}$$

We also assume that

$$K_M := \sup_{0<\epsilon\le 1} \sup_{\phi\in\mathcal{M}_1^+(\mathcal{P})} \sup_y \iint K_\epsilon(x,y,z)dx\,\phi(z)dz < +\infty. \tag{12.12}$$

Then, let $\rho_M := R_m^{-1}(K_M)$. The following boundedness result is straightforward:

**Proposition 12.1** (Upper bound for $\rho_\epsilon$). *Under assumptions (12.3), (12.11) and (12.12), all trajectories of (12.1) are forward-$\rho_M$-bounded from above in $\rho_\epsilon$, by which we mean that $\dot\rho_\epsilon(t) < 0$ as long as $\rho_\epsilon(t) > \rho_M$.*

Conversely, we can study conditions that ensure non-extinction of the population: $\rho_\epsilon(t) \ge \rho_m > 0$. For instance, let

$$\kappa_m(\rho) := \inf_{0<\epsilon\le 1} \inf_{\phi\in\mathcal{M}_+^1(\mathcal{P})} \inf_y \iint K_\epsilon(x,y,z)dx\,\phi(z)dz - R(y,\rho). \tag{12.13}$$

**Proposition 12.2** (Lower bound for $\rho_\epsilon$ under assumption (12.13)). *Under assumption (12.3) and if there exists $\rho_m > 0$ such that $\kappa_m(\rho_m) = 0$, with $\kappa_m$ defined in (12.13), then all trajectories of (12.1) are forward-$\rho_m$-bounded from below in $\rho_\epsilon$, by which we mean that $\dot\rho_\epsilon(t) > 0$ as long as $\rho_\epsilon(t) < \rho_m$.*

However, $\kappa_m(0) > 0$ is not expected to be a necessary condition. It is an open and challenging question to determine more general conditions for non-extinction, and study the set of extinction trajectories in cases when these conditions are not met.

For instance, let

$$\kappa_m'(\rho) := \inf_{0<\epsilon\le 1} \inf_{\phi\in\mathcal{M}_+^1(\mathcal{P})} \int \big( \iint K_\epsilon(x,y,z)dx\,\phi(z)dz - R(y,\rho)\big)\phi(y)dy. \tag{12.14}$$

**Proposition 12.3** (Lower bound for $\rho_\epsilon$ with a condition on (12.14)). *Under assumption (12.3) and if there exists $\rho_m > 0$ such that $\kappa_m'(\rho_m) = 0$ then all trajectories of (12.1) are forward-$\rho_m$-bounded from below in $\rho_\epsilon$.*

And likewise, let

$$\kappa_m'' := \inf_{0<\epsilon\le 1} \inf_{\phi\in\mathcal{M}_+^1(\mathcal{P})} \iiint K_\epsilon(x,y,z)\phi(y)\phi(z)\,dx\,dy\,dz. \tag{12.15}$$

Then, assuming also

$$\begin{aligned} &\exists R_M : \mathbb{R}_+ \to \mathbb{R}_+, \text{ increasing, with } R_M(0) \ge 0,\\ &R_M(+\infty) = +\infty \text{ and } \forall x, \quad R(x,\rho) \le R_M(\rho), \end{aligned} \tag{12.16}$$

we have $\dot\rho \ge \big(\kappa_m'' - R_M(\rho)\big)\rho$.

**Proposition 12.4** (Lower bound for $\rho_\epsilon$ with a condition on (12.15)). *Assume (12.16) holds and $\kappa_m'' > R_M(0)$ defined in (12.15). Then all trajectories of (12.1) are forward-$\rho_m$-bounded from below in $\rho_\epsilon$, with $\rho_m = R_M^{-1}(\kappa_m'') > 0$.*

## 12.3 The model without mutations

In order to see clearly the kind of results to be expected, we first study in detail a very simple example, which is equation (nM). The form of the birth rate assumes that the trait is perfectly transmitted from the females to their progeny, and the cross-fecundity between a male of trait $z$ and a female of trait $x$ depends only on the distance between $x$ and $z$ through $K_0$.

Assumptions (12.3) and (12.11) (for $R_m = \nu\rho$) obviously hold in case (nM). Assumption (12.12) holds with $K_M = \max_x K_0(x)$. However, $\kappa_m(\rho) = \inf K_0 - \rho$ so the non-extinction condition from Proposition 12.2 holds if and only if $\inf_x K_0(x) > 0$.

Even though $\inf_x K_0(x) = 0$, it seems standard to assume that $K_0$ is such that Proposition 12.4 holds, that is

$$\kappa_m'' = \inf_{\phi\in\mathcal{M}_+^1(\mathbb{R})} \int \big(K_0 * \phi\big)(x)\phi(x)\,dx > 0.$$

The same kind of assumption is used in [123] to prove an entropy-based stability result.

### 12.3.1 Proof of Theorem 12.1

In this section we prove Theorem 12.1. From equation (nM) we can compute

$$\epsilon\dot\rho_\epsilon(t) = \int n_\epsilon(t,x)\frac{K_0 * n_\epsilon(t,\,\cdot\,)}{\rho_\epsilon(t)}(x)\,dx - \nu\rho_\epsilon^2.$$

Assuming (12.6) yields $\frac{d}{dt}\int nK_0 * n = 2\int nK_0 * (\partial_t n)$. Hence

$$\epsilon\ddot\rho_\epsilon = -\nu\rho_\epsilon\dot\rho_\epsilon - \nu\rho_\epsilon\dot\rho_\epsilon - \frac{\dot\rho_\epsilon}{\rho_\epsilon^2}\int n_\epsilon K_0 * n_\epsilon + \frac{1}{2\rho_\epsilon}\frac{d}{dt}\int n_\epsilon K_0 * n_\epsilon + \frac{1}{\rho_\epsilon}\int \partial_t n_\epsilon K_0 * n_\epsilon.$$

We rewrite this as

$$\epsilon\ddot\rho_\epsilon = -\nu\rho_\epsilon\dot\rho_\epsilon - \frac{\dot\rho_\epsilon}{2\rho_\epsilon^2}\int n_\epsilon K_0 * n_\epsilon + \frac12\frac{d}{dt}\Big(\frac{1}{\rho_\epsilon}\int n_\epsilon K_0 * n_\epsilon - \nu\rho_\epsilon^2\Big) + \frac{1}{\epsilon}\int\Big(\frac{n_\epsilon(K_0 * n_\epsilon)^2}{\rho_\epsilon^2} - \nu n_\epsilon K_0 * n_\epsilon\Big),$$

and since $\epsilon\dot\rho_\epsilon = \frac{1}{\rho_\epsilon}\int n_\epsilon K_0 * n_\epsilon - \nu\rho_\epsilon^2$, we get

$$\frac{\epsilon}{2}\ddot\rho_\epsilon = -\frac{\dot\rho_\epsilon}{2\rho_\epsilon^2}\int n_\epsilon K_0 * n_\epsilon + \frac{1}{\epsilon}\int n_\epsilon\Big(\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\Big)^2. \tag{12.17}$$

From (12.17) a lot can be said. First, $\ddot\rho_\epsilon \geq -\frac{\dot\rho_\epsilon}{\epsilon\rho_\epsilon^2}\int n_\epsilon K_0 * n_\epsilon$, hence if $\dot\rho_\epsilon = 0$ then $\ddot\rho_\epsilon \geq 0$. In particular, $\rho_\epsilon$ has no strict local maximum. We can conclude that $\rho_\epsilon$ is either decreasing, increasing or decreasing-increasing, and since it is bounded, $\rho_\epsilon(t)$ must converge to some finite value $\rho_\epsilon^\infty$ as $t$ goes to $+\infty$.

Let $b_\epsilon(t) := \frac{1}{\rho_\epsilon^2(t)}\int n_\epsilon(t,x)(K_0 * n_\epsilon(t,\,\cdot\,))(x)dx \geq \kappa_m'' > 0$. Then from (12.17),

$$\frac{d}{dt}(\dot\rho_\epsilon)_- \leq -\frac{\kappa_m''}{\epsilon}(\dot\rho_\epsilon)_-.$$

Hence $(\dot\rho_\epsilon)_-(t) \leq e^{-\frac{\kappa_m'' t}{\epsilon}}(\dot\rho_\epsilon)_-(0)$. We write

$$\begin{aligned}\int_0^T |\dot\rho_\epsilon(t)|\,dt &\leq \int_0^T \dot\rho_\epsilon(t)\,dt + 2\int_0^T (\dot\rho_\epsilon)_-(t)dt\\ &\leq \rho_M + 2(\dot\rho_\epsilon)_-(0)\int_0^T e^{-\frac{\kappa_m'' t}{\epsilon}}dt\\ &\leq \rho_M + 2(\dot\rho_\epsilon)_-(0)\frac{\epsilon}{\kappa_m''}\big(1 - e^{-\frac{\kappa_m'' T}{\epsilon}}\big).\end{aligned}$$

Therefore, under the mild assumption (12.5), the family $(\rho_\epsilon)_\epsilon$ is uniformly bounded in $BV(\mathbb{R}_+)$.

We now establish equation (12.7). Going back to equation (12.17), and integrating this one over $[0,T]$ for $T>0$, we obtain

$$\int_0^T \int_{\mathbb{R}} n_\epsilon \Big(\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\Big)^2 dx\,dt = \epsilon \int_0^T \frac{\dot\rho_\epsilon}{2\rho^2} \int n_\epsilon K_0 * n_\epsilon\,dx\,dt + \frac{\epsilon^2}{2}(\dot\rho_\epsilon(T) - \dot\rho_\epsilon(0)). \tag{12.18}$$

Since $\rho_\epsilon$ is locally $BV$ uniformly in $\epsilon$ and using (12.5) and (12.6), we deduce that

$$\int_0^T \int_{\mathbb{R}} n_\epsilon \Big(\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\Big)^2 dx\,dt = O(\epsilon),$$

which is equation (12.7).

### 12.3.2 Concentration of Dirac masses

Formally, in the limit $\epsilon \to 0$ the previous estimation yields

$$\int n(t,x)\Big(\frac{K_0 * n(t,\,\cdot\,)}{\rho(t)}(x) - \nu\rho(t)\Big)^2 dx = 0. \tag{12.19}$$

It turns out that combinations of Dirac masses are admissible solutions to (12.19), $n = \sum_{i=1}^N \rho_i \delta_{x_i}$, $\rho_i > 0$ with $\sum_{i=1}^N \rho_i = \rho$, and

$$\sum_{i=1}^N \rho_i \Big(\sum_{j=1}^N \frac{\rho_j}{\rho} K_0(x_i - x_j) - \nu\rho\Big)^2 = 0,$$

so for all $i \in \{1,\dots,N\}$,

$$\sum_{j=1}^N \frac{\rho_j}{\rho} K_0(x_i - x_j) = \nu\rho. \tag{12.20}$$

We define the matrix $\underline{\mathbf{K}}$, whose coefficient with indices $(i,j)$ is equal to $K_0(x_i - x_j)$. $\underline{\mathbf{K}}$ is symmetric with positive coefficients and constant main diagonal (equal to $K_0(0)$). If the family $(x_i)_{1\le i\le N}$ is given, the problem amounts to finding a positive vector $P$ such that $\underline{\mathbf{K}}P = \mathbb{1}$. (Then $\rho = 1/\mathbb{1}^T P$ and $\rho_i = P_i\rho^2$).

It is easily checked that if

$$\max_{i\neq j} K_0(x_i - x_j) < \frac{K_0(0)}{N-1}$$

then $\underline{\mathbf{K}}$ is invertible (in this case indeed, $\underline{\mathbf{K}}$ is strictly diagonally dominant). It is worth noting that if $N=2$ and $\max K_0 = K_0(0)$, with the maximum of $K_0$ being reached only at 0, then this is always satisfied. However, in the generic case when $(x_i)_i$ is such that $\underline{\mathbf{K}}$ is invertible, it remains unclear whether $P := \underline{\mathbf{K}}^{-1}\mathbb{1} > 0$ or not.

In spite of this, an alternative viewpoint using a Lyapunov functional helps describing the asymptotically stable solutions, as detailed below.

### 12.3.3 A Lyapunov concentration result

Let $n^0 \in \mathcal{M}_+(\mathcal{P})$ and $q^0 = n^0/\int_X n^0 \in \mathcal{M}_+^1(\mathcal{P})$. We introduce $K_S : \mathcal{P}^2 \to \mathbb{R}_+$, $R_0 : \mathcal{P} \to \mathbb{R}_+$ and $R_1 : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying the following assumptions:

$$K_S \in \mathcal{C}_b(\mathcal{P}^2, \mathbb{R}_+) \text{ is symmetric: } \forall x,y \in \mathcal{P},\ K_S(x,y) = K_S(y,x), \tag{$K_S$S}$$

$$\forall \xi \in \mathcal{M}(\mathcal{P})\backslash\{0\}, \quad \iint K_S(x,y)\xi(x)\xi(y)dxdy > 0, \tag{$K_S$P}$$

$$supp(q^0) \text{ is compact or } R_0 \text{ is proper}, \tag{$R_0$P}$$

$$R_1 \text{ is increasing and proper}, \tag{$R_1$P}$$

$$\exists! x_M \in supp(q^0), \quad y \mapsto K_S(x_M,y) - R_0(y) \text{ reaches its maximum at } x_M. \tag{Max}$$

Then, we consider the special form of equation (12.1) given by (gnM). Defining $q(t,x) := n_\epsilon(\epsilon t, x)/\rho_\epsilon(t)$ in this case, we obtain

$$\partial_t q(t,x) = q(t,x)\Big(\int K_S(x,y)q(t,y)dy - R_0(x)\Big) \\ - q(t,x)\int q(t,x')\Big(\int K_S(x',y)q(t,y)dy - R_0(x')\Big)dx', \quad q(0,x) = q^0(x). \tag{12.21}$$

Then we simply need to study the asymptotic behavior of $q$ as $t \to +\infty$ to be able to describe that of $n_\epsilon$ as $\epsilon \to 0$. Thanks to the structure of (12.21) we have:

**Theorem 12.5.** *Under assumptions* ($K_S$S),($K_S$P),($R_0$P),($R_1$P) *and* (Max), $\delta_{x_M}$ *is asymptotically stable for* (12.21).

The proof relies on Proposition 4.9 (stated in Section 4.4.2, where relevant definitions for Lyapunov functions and a proof are given). We get a Lyapunov functional for (12.21) by defining

$$J(q) := \frac{1}{2}\iint K_S(x,y)q(x)q(y)dxdy - \int R_0(x)q(x)dx. \tag{12.22}$$

Indeed, along an orbit of (12.21) we have

$$\frac{d}{dt}J(q(t,\cdot)) = \int q(t,x)\Big(\int K_S(x,y)q(t,y)dy - R_0(x)\Big)^2 dx \\ - \Big(\int q(t,x)\Big(\int K_S(x,y)q(t,y)dy - R_0(x)\Big)dx\Big)^2 \geq 0,$$

with equality (by Cauchy-Schwarz inequality) if and only if

$$\int K_S(x,y)q(t,y)dy - R_0(x) \equiv C \in \mathbb{R} \text{ on } supp(q(t,\cdot)),$$

so that we have strict monotonicity except if $q(t,\cdot)$ is a rest point for the dynamics of (12.21). This Lyapunov functional can be seen as an embodiment of the "positive correlation" property from game dynamics (see [116]), and this feature has been exploited to get a gradient flow formulation of a non-local model with a diffusion term in [122], where the kernel acts for death induced by competition rather than for birth as is the case here.

In order to use the Lyapunov functional properly, we need $\{q(t,\cdot),\, t \geq 0\} \subset \mathcal{M}_+^1(\mathcal{P})$ to be relatively compact for a topology for which $J$ is continuous and Fréchet-differentiable. This is the case if either $q^0$ is compactly supported in $\mathcal{P}$, or if $K$ is bounded and $R_0$ is proper (by Prohorov's theorem), for the weak* topology on $\mathcal{M}_+^1(\mathcal{P})$.

The study of the maximizer sets for $J$ is greatly simplified by ($K_S$P):

**Lemma 12.1.** *Under* ($K_S$P), *the functional $J$ is strictly convex on the convex set $\mathcal{M}_+^1(\mathcal{P})$.*

*Therefore its local maximum points are extreme points of $\mathcal{M}_+^1(\mathcal{P})$, that is Dirac masses. The Dirac mass $\delta_x$ is a local maximizer of $J$ only if $y \mapsto K_S(x,y) - R_0(y)$ reaches its maximum at $x$.*

*Proof.* For $q_1, q_2 \in \mathcal{M}_+^1(\mathcal{P})$ and $\theta \in [0,1]$ we compute

$$J(\theta q_1 + (1-\theta)q_2) = \iint K_S\big(\theta^2 q_1 q_1 + (1-\theta)^2 q_2 q_2 + 2\theta(1-\theta)q_1 q_2\big) - \int R_0\big(\theta q_1 + (1-\theta)q_2\big) \\ = \theta J(q_1) + (1-\theta)J(q_2) - \theta(1-\theta)\iint K_S(x,y)(q_1-q_2)(x)(q_1-q_2)(y)dxdy.$$

Therefore ($K_S$P) (with $\xi = q_1 - q_2$) implies that $J$ is strictly convex.

If $J$ reaches a local maximum at $\xi \in \mathcal{M}_+^1(\mathcal{P})$ belonging to some interval $(\xi_-, \xi_+)$, that is $\xi = \theta\xi_- + (1-\theta)\xi_+$ for some $\theta \in (0,1)$ with $\xi_\pm \in \mathcal{M}_+^1(\mathcal{P})$, then for $\epsilon > 0$ small enough we have

$$J(\xi) < \frac{1}{2}\big(J(\xi + \epsilon(\xi_+ - \xi_-)) + J(\xi - \epsilon(\xi_+ - \xi_-))\big) \leq J(\xi),$$

where the left inequality holds by strict convexity and the right one by the local maximum condition. This is absurd, hence local maxima are only reached at extreme points.

The support of an extreme probability measure must be reduced to a singleton: otherwise, we can construct a segment on which the measure lies by exchanging mass between any two separable points of the support. Conversely, a Dirac mass is obviously extreme, as any segment to which it belongs would consist of probability measures with the same support, reduced to a singleton.

Then, the first-order optimality condition for $J$ at $\delta_x$ reads: for all admissible perturbation $h$,

$$\int \big(K(x,y) - R_0(y)\big)h(y)dy \leq 0,$$

and admissible perturbations have the general form $h = -\delta_x + h_0$, with $h_0 \in \mathcal{M}_+^1(\mathcal{P})$, whence the last point.

□

Thanks to (Max) we get that $\{\delta_{x_M}\}$ is a local maximizer set of $J$ for which $J$ is a strict Lyapunov function (and that there is no other local maximizer set of $J$). By Proposition 4.9, it is asymptotically stable.

## 12.4 $BV$ estimates on the total population

In this section, we derive BV estimates assuming that $R$ is independent from the trait variable and features a linear dependency on $\rho$, which is specified by assumption (12.4). Thereafter we address the difficulties encountered when $R$ has a general form.

### 12.4.1 Linear dependency on the competition variable in the AF model

Although the asymptotic behavior of $n_\epsilon$ solution to (12.1) may be difficult to obtain in general, under some assumptions on $K$ and $R$, the total population $\rho_\epsilon$ can be proved to have bounded variations.

Recall that, integrating equation (12.1), we have

$$\epsilon\dot\rho_\epsilon = \frac{1}{\rho_\epsilon}\iiint K_\epsilon(x,y,z)n_\epsilon(t,y)n_\epsilon(t,z)dx\,dy\,dz - \int R(x,\rho_\epsilon)n_\epsilon(t,x)\,dx.$$

The proofs of Theorems 12.2 and 12.3 rely on estimates obtained through the equation satisfied by $\ddot\rho_\epsilon$. In general, we start from

$$\begin{aligned}\epsilon\ddot\rho_\epsilon = &-\frac{\dot\rho_\epsilon}{\rho_\epsilon^2}\iiint K_\epsilon(x,y,z)n_\epsilon(t,y)n_\epsilon(t,z)dx\,dy\,dz\\ &+\frac{1}{\rho_\epsilon}\iiint K_\epsilon(x,y,z)\big(\partial_t n_\epsilon(t,y)n_\epsilon(t,z) + n_\epsilon(t,y)\partial_t n_\epsilon(t,z)\big)dx\,dy\,dz\\ &-\dot\rho_\epsilon\int \partial_\rho R(x,\rho)n_\epsilon(t,x)dx\\ &-\int R(x,\rho_\epsilon)\Big(\frac{1}{\epsilon\rho_\epsilon}\iint K_\epsilon(x,y,z)n_\epsilon(t,y)n_\epsilon(t,z)dy\,dz - R(x,\rho_\epsilon)n_\epsilon(t,x)\Big)dx. \quad (12.23)\end{aligned}$$

*Proof of Theorem 12.2.* We treat the case of the model with asymmetric fecundity. Then, $\rho_\epsilon$ satisfies

$$\epsilon\dot\rho_\epsilon = \int B(x)n_\epsilon(t,x)dx - \nu\rho_\epsilon^2,$$

and (12.23) reads

$$\begin{aligned}\epsilon\ddot\rho &= \int B(x)\partial_t n_\epsilon(t,x)dx - 2\nu\rho_\epsilon\dot\rho_\epsilon\\ &= -\nu\rho_\epsilon\dot\rho_\epsilon + \frac{\nu^2}{\epsilon}\rho_\epsilon^3 - \frac{\nu\rho_\epsilon}{\epsilon}\int B(x)n_\epsilon(t,x)dx\\ &+\frac{1}{\epsilon\rho_\epsilon}\iiint \alpha_\epsilon(x,y,z)B(x)B(y)n_\epsilon(t,y)n_\epsilon(t,z)dx\,dy\,dz - \frac{\nu\rho_\epsilon}{\epsilon}\int B(x)n_\epsilon(t,x)dx.\end{aligned}$$

Which we rewrite as

$$\epsilon\frac{d}{dt}\dot\rho_\epsilon = -\nu\rho_\epsilon\dot\rho_\epsilon + \overbrace{\frac{\rho_\epsilon}{\epsilon}\left(\frac{\int B(x)n_\epsilon(t,x)dx}{\rho_\epsilon} - \nu\rho_\epsilon\right)^2}^{\text{demographic stabilization}} + \underbrace{\frac{1}{\epsilon\rho_\epsilon}\Big(\iiint \alpha_\epsilon(x,y,z)B(x)B(y)n_\epsilon(t,y)n_\epsilon(t,z)dx\,dy\,dz - \Big(\int B(x)n_\epsilon(t,x)dx\Big)^2\Big)}_{\text{mixing-induced fecundity variation}}. \quad (12.24)$$

In order to apply the same technique as for the simple case (nM), we need to assume that the mixing-induced fecundity variation term is bounded from below.

Under (12.8), we obtain from (12.24) and Proposition 12.1

$$\epsilon\frac{d}{dt}\dot\rho_\epsilon \geq -\nu\rho_\epsilon\dot\rho_\epsilon - C. \quad (12.25)$$

And from Proposition 12.4, we deduce

$$\frac{d}{dt}(\dot\rho_\epsilon)_- \leq -\frac{\nu\rho_m}{\epsilon}(\dot\rho_\epsilon)_- + \frac{C}{\epsilon},$$

and thus $(\dot\rho_\epsilon)_-(t) \leq e^{-\frac{\nu\rho_m t}{\epsilon}}(\dot\rho_\epsilon)_-(0) + \frac{C}{\nu\rho_m}\big(1-e^{-\frac{\nu\rho_m t}{\epsilon}}\big)$. Then we use the same argument we used to treat case without mutations in the previous section, which proves uniform boundedness of $(\rho_\epsilon)_\epsilon$ in $BV(0,T)$, for all $T>0$. □

We discuss assumption (12.8). First, if $B$ is constant then it is obviously satisfied. Secondly, by taking $\phi$ concentrated at a point $x_M$ where $B$ reaches its maximum, we obtain

$$\int \alpha_\epsilon(x,x_M,x_M)B(x)dx \geq B(x_M) - C\epsilon.$$

Recalling that $\int \alpha_\epsilon(x,y,z)dx = 1$ for all $y,z$, this implies that as $\epsilon$ goes to 0, $\alpha_\epsilon(\,\cdot\,,x_M,x_M)$ is concentrated at points where $B$ is equal to its maximum $B(x_M)$, which is a restrictive necessary condition for (12.8) to hold.

Thirdly, we state a sufficient condition: if $\alpha_\epsilon(\,\cdot\,,y,z) \to \alpha_0(y,z) \in \mathcal{M}_+^1(\mathcal{P})$ with either

$$\forall y,z, \quad \int \alpha_0(y,z)(x)B(x)dx \geq B(y)$$

or

$$\forall y,z, \quad \int \alpha_0(y,z)(x)B(x)dx \geq B(z),$$

and if convergence is sufficiently fast, then (12.8) holds. In the first case this is a consequence of the Cauchy-Schwarz inequality, and in the second case we simply obtain that the left-hand side in (12.8) converges to 0 as $\epsilon\to 0$. In particular, we may assume $\alpha_\epsilon(x,y,z) = \frac{1}{\epsilon}G\big(\frac{x-y}{\epsilon}\big)$ or $\frac{1}{\epsilon}G\big(\frac{x-z}{\epsilon}\big)$ for some appropriate kernel $G$. These situations are those we have in mind, although (12.8) in all generality may allow for some other cases.

All in all, (12.8) means that the trait inheritance pattern $\alpha_\epsilon$ does not allow next-generation's fecundity to get smaller than the current one's as $\epsilon\to 0$. Unsurprisingly, this dissipative feature implies that the variations of $\rho$ can be controlled as $\epsilon\to 0$, as stated in Theorem 12.2.

### 12.4.2 Linear dependency on the competition variable in the ATH model

We now address the case of the model with asymmetric trait heredity, which we refer to as the ATH model.

**Remark 12.1.** *In order to apply the same technique as for the model without mutations addressed in Section 12.3, we need a convergence assumption on $G_\epsilon$ as $\epsilon$ goes to 0. Specifically, we use the following assumption: for all Lipschitz function $\phi$, we have*

$$G_\epsilon * \phi = \phi + O(\epsilon). \quad (12.26)$$

*This assumption on the convergence of $G_\epsilon$ as $\epsilon$ goes to 0 holds in the typical example where $G_\epsilon$ is Gaussian with variance $\epsilon^2$. It means that there exists $C \in \mathbb{R}_+^*$ such that for all $\epsilon > 0$, $\phi \in W^{1,\infty}$ with $\|\phi\|_{Lip} \le 1$ and $\psi \in L^\infty$ with $\|\psi\|_{L^\infty} \le 1$,*

$$\left| \int \psi(x)(G_\epsilon * \phi)(x)dx - \int \psi(x)\phi(x)dx \right| \le C\epsilon.$$

*Specifically, we write $G_\epsilon(x) = \frac{1}{(2\pi\epsilon^2)^{d/2}} e^{-x^2/2\epsilon^2}$. Then we compute*

$$\begin{aligned}\delta &:= \left| \int \psi(x)(G_\epsilon * \phi)(x)dx - \int \psi(x)\phi(x)dx \right| \\ &\le \int |\psi(x)||G_\epsilon * \phi(x) - \phi(x)|dx \le \int\int \frac{1}{(2\pi\epsilon^2)^{d/2}} e^{-\frac{(x-y)^2}{2\epsilon^2}} |\phi(y) - \phi(x)|dydx.\end{aligned}$$

*We apply the change of variables $\hat{y} = (2\epsilon)^{-1}(y - x)$, so $d\hat{y} = (2\epsilon)^{-d}dy$, to get*

$$\delta \le \pi^{-d/2} \int\int e^{-\hat{y}^2} |\phi(x + 2\epsilon\hat{y}) - \phi(x)|d\hat{y}dx \le \frac{2\|\phi\|_{Lip}}{(2\pi)^{d/2}}\epsilon.$$

*Proof of Theorem 12.3.* Departing from (ATH), the equation satisfied by $\rho_\epsilon$ reads

$$\epsilon \frac{d}{dt}\rho_\epsilon(t) = \int_{\mathbb{R}} \left( \frac{1}{\rho_\epsilon(t)} K * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) - \nu \rho_\epsilon(t) n_\epsilon(t, x) \right) dx.$$

Differentiating this equation, we obtain

$$\begin{aligned}\epsilon\ddot{\rho}_\epsilon(t) = \quad & \frac{1}{\rho_\epsilon(t)} \int_{\mathbb{R}} \left[ K_0 * \partial_t n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) + K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * \partial_t n_\epsilon(t, \cdot)(x) \right] dx \\ & - \nu\rho_\epsilon(t) \frac{d}{dt}\rho_\epsilon(t) - \nu \int \partial_t n_\epsilon(t, x)\rho_\epsilon(t)dx \\ & - \frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)} \int \left[ K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) \right] dx\end{aligned}$$

By the same trick as in Section 12.3, assuming (12.6) induces

$$\begin{aligned}\epsilon\ddot{\rho}_\epsilon(t) = \quad & \frac{1}{2\rho_\epsilon(t)} \frac{d}{dt} \left[ \int K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x)dx \right] \\ & + \frac{1}{\rho_\epsilon(t)} \int \left[ G_\epsilon * (K_0 * n_\epsilon(t, \cdot))(x) \partial_t n_\epsilon(t, x) \right] dx \\ & - \nu\rho_\epsilon(t) \frac{d}{dt}\rho_\epsilon(t) - \nu \int \partial_t n_\epsilon(t, x)\rho_\epsilon(t)dx \\ & - \frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)} \int \left[ K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) \right] dx.\end{aligned}$$

Then we compute

$$\begin{aligned}\epsilon\ddot{\rho}_\epsilon(t) = & \frac{1}{2\rho_\epsilon(t)} \frac{d}{dt} \left[ \int K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x)dx \right] \\ & + \frac{1}{\epsilon} \frac{1}{\rho_\epsilon(t)} \int G_\epsilon * (K_0 * n_\epsilon(t, \cdot))(x) \left[ \frac{1}{\rho_\epsilon(t)} K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) - \nu n_\epsilon(t, x)\rho_\epsilon(t) \right] dx \\ & - \nu\rho_\epsilon(t) \frac{d}{dt}\rho_\epsilon(t) - \frac{\nu}{\epsilon}\rho_\epsilon(t) \int \left[ \frac{1}{\rho_\epsilon(t)} K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) - \nu n_\epsilon(t, x)\rho_\epsilon(t) \right] dx \\ & - \frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)} \int \left[ K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) \right] dx,\end{aligned}$$

and get

$$\begin{aligned}\epsilon\ddot{\rho}_\epsilon(t) = & -\nu\rho_\epsilon(t) \frac{d}{dt}\rho_\epsilon(t) + \frac{1}{2} \frac{d}{dt} \left[ \int \frac{1}{\rho_\epsilon(t)} K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x)dx \right] \\ & - \frac{1}{2} \frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)} \int \left[ K_0 * n_\epsilon(t, \cdot)(x) G_\epsilon * n_\epsilon(t, \cdot)(x) \right] dx \\ & + \frac{1}{\epsilon} \int (G_\epsilon * n_\epsilon) \left[ G_\epsilon * (K_0 * n_\epsilon) \frac{(K_0 * n_\epsilon)}{\rho_\epsilon^2} - 2\nu K_0 * n \right] dx + \frac{1}{\epsilon}\nu^2\rho_\epsilon^2 \int G_\epsilon * n_\epsilon \, dx.\end{aligned}$$

We rewrite this as

$$\epsilon\ddot{\rho}_\epsilon(t) = -\nu\rho_\epsilon(t)\dot{\rho}_\epsilon(t) - \frac{1}{2}\frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)}\int K_0 * n_\epsilon(t,\cdot)G_\epsilon * n_\epsilon(t,\cdot) + \frac{1}{\epsilon}\int (G_\epsilon * n_\epsilon)\left[\frac{(K_0 * n_\epsilon)}{\rho_\epsilon} - \nu\rho_\epsilon\right]^2 dx + \frac{1}{\epsilon\rho_\epsilon^2(t)}\int (K_0 * n_\epsilon)(G_\epsilon * n_\epsilon)\Big(G_\epsilon * (K_0 * n_\epsilon) - K_0 * n_\epsilon\Big)dx.$$

Now we use the convergence assumption (12.26) on $G_\epsilon$. We simply need to check that $\phi(x) := \int K_0(x-y)n_\epsilon(t,y)dy$ is Lipschitz. But obviously, $|\phi'| \le \|K_0'\|_{L^\infty}\rho_\epsilon(t)$. Hence

$$\frac{\epsilon}{2}\ddot{\rho}_\epsilon(t) = -\nu\rho_\epsilon(t)\dot{\rho}_\epsilon(t) - \frac{1}{2}\frac{\dot{\rho}_\epsilon(t)}{\rho_\epsilon^2(t)}\int K_0 * n_\epsilon(t,\cdot)G_\epsilon * n_\epsilon(t,\cdot) + \frac{1}{\epsilon}\int (G_\epsilon * n_\epsilon)\left[\frac{(K_0 * n_\epsilon)}{\rho_\epsilon} - \nu\rho_\epsilon\right]^2 dx + O(1). \tag{12.27}$$

Thanks to (12.10), we obtain

$$\frac{\epsilon}{2}\frac{d}{dt}(\dot{\rho}_\epsilon(t))_- \le -\big(\frac{1}{2}\eta_0 + \nu\rho_\epsilon(t)\big)(\dot{\rho}_\epsilon(t))_- + O(1).$$

Then, $\rho_\epsilon$ is bounded in $BV_{\rm loc}(\mathbb{R}_+)$ uniformly in $\epsilon$. Indeed, we obtain that for some constants $C_1, C_2 > 0$,

$$(\dot{\rho}_\epsilon(t))_- \le e^{-C_1 t/\epsilon}\Big((\dot{\rho}_\epsilon(0))_- + \frac{C_2}{\epsilon}\int_0^t e^{C_1 t'/\epsilon}dt'\Big),$$

hence

$$(\dot{\rho}_\epsilon(t))_- \le (\dot{\rho}_\epsilon(0))_- e^{-C_1 t/\epsilon} + \frac{C_2}{C_1}\big(1 - e^{-C_1 t/\epsilon}\big).$$

As in the proof of Theorem 12.1, we deduce that for all $T > 0$, $(\rho_\epsilon)_\epsilon$ is uniformly (in $\epsilon$) bounded in $BV([0,T])$.

Going back to (12.27), we deduce the estimate, for $T > 0$

$$\int_0^T\int_x (G_\epsilon * n_\epsilon)\left[\frac{K_0 * n_\epsilon}{\rho_\epsilon} - \nu\rho_\epsilon\right]^2 dxdt = O(\epsilon),$$

as in the proof of Theorem 12.1.

□

### 12.4.3 Questions and difficulties for the general case

Firstly, we address the case of a general saturation term for the AF model, featuring the competition effect and the trait-dependency:

$$R \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}_+; \mathbb{R}_+), \quad K(x,y,z) = B(y)\alpha_\epsilon(x,y,z), \quad \forall y,z, \int \alpha_\epsilon(x,y,z)dx = 1.$$

Then, we find

$$\epsilon\frac{d}{dt}\dot{\rho}_\epsilon = \int \big(B(x) - R(x,\rho_\epsilon)\big)\partial_t n_\epsilon(t,x)dx - \dot{\rho}_\epsilon\int \partial_\rho R(x,\rho_\epsilon)n_\epsilon(t,x)dx$$
$$= -\dot{\rho}_\epsilon\int \partial_\rho R(x,\rho_\epsilon)n_\epsilon(t,x)dx + \frac{1}{\epsilon}\int n_\epsilon(t,x)\big(R(x,\rho_\epsilon) - B(x)\big)^2$$
$$+ \frac{1}{\epsilon}\int \big(B(x) - R(x,\rho_\epsilon)\big)\Big(\frac{1}{\rho_\epsilon}\iint \alpha_\epsilon(x,y,z)B(y)n_\epsilon(t,y)n_\epsilon(t,z)dy\,dz - B(x)n_\epsilon(t,x)\Big)dx$$

The last term can be seen as the integral of the net fitness $B - R(\cdot\,,\rho_\epsilon)$ weighted by a fecundity variation $\Delta_{n_\epsilon(t,\cdot)}B$ (with $\int \Delta_{n_\epsilon(t,\cdot)}B(x)dx = 0$).

To apply the same argument as before, we need to assume

$$\begin{aligned} &\exists C>0,\ \forall \epsilon>0,\ \forall y,z,\ \forall \phi\in L^1_+ \text{ with } \|\phi\|_{L^1}=1,\\ &\Big\| \iint \alpha_\epsilon(\,\cdot\,,y,z)B(y)\phi(y)\phi(z)dydz - B(\,\cdot\,)\phi(\,\cdot\,)\Big\|_{L^1}\le C\epsilon, \end{aligned} \tag{12.28}$$

and we also assume that

$$\forall \rho\le \rho_M, C_f(\rho):=\|B(\,\cdot\,)-R(\,\cdot\,,\rho)\|_\infty<\infty,\quad \overline{C_f}=\sup_{0\le\rho\le\rho_M} C_f(\rho). \tag{12.29}$$

Under assumptions (12.28) and (12.29), this additional term is treated as in the case (12.4), replacing the negative constant on the right-hand side of (12.25) by $-\rho_M C\overline{C_f}$, which gives

$$\epsilon\frac{d}{dt}\dot\rho_\epsilon(t)\ge -\dot\rho_\epsilon\int \partial_\rho R(x,\rho_\epsilon)n_\epsilon(t,x)dx-\rho_M C\overline{C_f}.$$

Therefore, similarly to the proof of Theorem 12.2 we obtain

**Lemma 12.2.** *Assume (12.28) and (12.29). Then, for all $T>0$, $(\rho_\epsilon)_\epsilon$ is uniformly in $\epsilon$ bounded in $BV([0,T])$.*

Secondly, we address the case of a general death term for the ATH model:

$$R\in\mathcal{C}^1(\mathbb{R}^d\times\mathbb{R}_+;\mathbb{R}_+),\quad K_\epsilon(x,y,z)=G_\epsilon(x-z)K_0(x-y).$$

To see clearly where the difficulty lies, we replace $G_\epsilon(x-z)$ by $\delta_{x=z}$ (letting $\epsilon\to 0$ in this term only). For simplicity, we define

$$\zeta(t,x):=\frac{K_0*n_\epsilon(t,\,\cdot\,)}{\rho_\epsilon(t)},\quad Q(t):=\int\partial_\rho R(x,\rho_\epsilon(t))n_\epsilon(t,x)dx.$$

After computations similar to the previous ones, we find

$$\frac12\epsilon\frac{d}{dt}\dot\rho_\epsilon=-\frac{\dot\rho_\epsilon}{2\rho_\epsilon}\int n_\epsilon\zeta+\frac1\epsilon\int n_\epsilon\Big[\zeta^2-R\zeta+\frac{R+Q}{2}\big(R-\zeta\big)\Big], \tag{12.30}$$

and the term in $\frac1\epsilon$ rewrites

$$\int n(\zeta-\frac{R+Q}{2})(\zeta-R).$$

Meanwhile, one can check that

$$\epsilon\dot\rho_\epsilon=\int n_\epsilon(\zeta-R).$$

When $\dot\rho_\epsilon\le 0$ we would like to prove that the term in $\frac1\epsilon$ in (12.30) is non-negative. We could be less restrictive and simply require $\ddot\rho_\epsilon\ge 0$. This reads (with $q_\epsilon(t,x)=n_\epsilon(t,x)/\rho_\epsilon(t)$):

$$\int q_\epsilon(t,x)\big(\zeta(t,x)-R(x,\rho_\epsilon(t))\big)\Big(\zeta(t,x)-\frac{R(x,\rho_\epsilon(t))+Q(t)}{2}-\int q_\epsilon(t,y)\zeta(t,y)dy\Big)dx\ge 0$$

if

$$\int q_\epsilon(t,x)\big(\zeta(t,x)-R(x,\rho_\epsilon(t))\big)dx\le 0.$$

We do not treat the general case but by a straightforward computation we can deduce

**Lemma 12.3.** *If $R(x,\rho)=R_1(\rho)$ and $\rho R_1'(\rho)\ge R_1(\rho)$, then $\dot\rho_\epsilon\le 0$ implies $\ddot\rho_\epsilon\ge-\frac{\dot\rho_\epsilon}{2\rho_\epsilon}\int n\zeta$. Then in particular for all $T>0$, $(\rho_\epsilon)_\epsilon$ is uniformly in $\epsilon$ bounded in $BV([0,T])$.*

For instance, this is the case if $R_1(\rho)=\nu\rho^\gamma$ for some $\gamma\ge 1$.

## 12.5 Lyapunov approach

As in Section 12.3.3, we define $q_\epsilon(t,x) := n_\epsilon(\epsilon t, x)/\rho_\epsilon(t)$. In the general case (12.1) under the single assumption $R(x,\rho) = R_0(x) + R_1(\rho)$ we derive the following equation:

$$\begin{cases} \partial_t q_\epsilon(t,x) = \iint K_\epsilon(x,y,z) q_\epsilon(t,y) q_\epsilon(t,z) dy dz - R_0(x) q_\epsilon(t,x) \\ -q_\epsilon(t,x)\Big( \iiint K_\epsilon(x',y,z) q_\epsilon(t,y) q_\epsilon(t,z) dx' dy dz - \int R_0(x') q_\epsilon(t,x') dx' \Big), \\ q_\epsilon(0,x) = q_\epsilon^0(x). \end{cases} \tag{12.31}$$

A natural candidate Lyapunov functional is given by

$$J^\epsilon(q) := \frac{1}{2} \iiint K_\epsilon^S(x,y,z) q(y) q(z) dx dy dz - \int R_0(x) q(x) dx,$$

where $K_\epsilon^S$ is the symmetrization of $K_\epsilon$:

$$K_\epsilon^S(x,y,z) = \frac{K_\epsilon(x,z,y) + K_\epsilon(x,y,z)}{2}.$$

We can compute along an orbit of (12.31):

$$\begin{aligned} \frac{d}{dt} J^\epsilon(q_\epsilon(t,\cdot)) &= \int \partial_t q_\epsilon(t,y) \Big( \iint K_\epsilon^S(x,y,z) q_\epsilon(t,z) dz dx - R_0(y) \Big) dy, \\ &= \int \cdots \int K_\epsilon^S(x,y,z) K_\epsilon^S(y,y',z') q_\epsilon(t,y') q_\epsilon(t,z') q_\epsilon(t,z) dz' dy' dz dy dx \\ &- \iiint K_\epsilon^S(x,y,z) \big( R_0(x) + R_0(y) \big) q_\epsilon(t,y) q_\epsilon(t,z) dx dy dz + \int q_\epsilon(t,x) R_0^2(x) dx \\ &- \Big( \iiint K_\epsilon^S(x,y,z) q_\epsilon(t,y) q_\epsilon(t,z) dx dy dz - \int q_\epsilon(t,y) R_0(y) dy \Big)^2. \end{aligned}$$

First, in the special case $R_0 \equiv 0$, to get a non-decreasing $J^\epsilon$ along orbits we need to assume

$$\forall \xi \in \mathcal{M}_+^1(\mathcal{P}), \quad \int \cdots \int K_\epsilon^S(x,y,z) K_\epsilon^S(y,y',z') \xi(y') \xi(z') \xi(z) dz' dy' dz dy dx \geq \Big( \iiint K_\epsilon^S(x,y,z) \xi(y) \xi(z) dx dy dz \Big)^2, \tag{12.32}$$

which should be interpreted as an increase of fecundity from parents to offspring, with equality only if the dynamic is at rest, that is

$$\iint K_\epsilon^S(\,\cdot\,,y,z) \xi(y) \xi(z) dy dz \text{ is constant on } supp(\xi).$$

In other words, to get a Lyapunov functional requires a perfect analogue of Cauchy-Schwarz inequality.

Another case where this approach seems to yield some results is for (AF) with constant $B$, that is under the assumption

$$\exists B > 0,\ \forall \epsilon > 0,\ \forall y, z, \quad \int K_\epsilon(x,y,z) dx = B.$$

In this case we write $K_\epsilon = B\alpha_\epsilon$ and get $J^\epsilon(q) = \frac{B}{2} - \int q(y) R_0(y) dy$ so that

$$\begin{aligned} \frac{d}{dt} J^\epsilon(q_\epsilon(t,\cdot)) = \int q_\epsilon(t,y) R_0^2(y) dy - \Big( \int q_\epsilon(t,y) R_0(y) dy \Big)^2 \\ + B\Big( \int q_\epsilon(t,y) R_0(y) dy - \iiint R_0(x) \alpha_\epsilon(x,y,z) q_\epsilon(t,y) q_\epsilon(t,z) dx dy dz \Big). \end{aligned}$$

One possible additional assumption is therefore

$$\forall \xi \in \mathcal{M}_+^1(\mathcal{P}), \quad \int R_0(y) \xi(y) dy \geq \iiint R_0(x) \alpha_\epsilon(x,y,z) \xi(y) \xi(z) dx dy dz, \tag{12.33}$$

which should be interpreted as a decrease of the death rate from parents to offspring.

In either one of the two above cases ($R_0 = 0$ and (12.32) or $K_\epsilon = B\alpha_\epsilon$ and (12.33)), and in the intermediate case when we assume a more complex inequality involving $K_\epsilon$ and $R$, the consequence of the assumptions is that $J^\epsilon$ is a Lyapunov function indeed for the dynamics of (12.31). As in Section 12.3.3, under additional assumptions to get orbit relative compactness (proper death rate yielding uniform tightness in the sense of Section 4.4.2), and functional convexity, we could obtain concentration to Dirac masses extending Theorem 12.5.

In fact, the more realistic assumptions such as (12.8), (12.28) or (12.26) do not imply that $J^\epsilon$ itself is a Lyapunov function, but rather that along an orbit of (12.31),

$$\frac{d}{dt}J^\epsilon(q_\epsilon(t,\cdot)) = j^0(q_\epsilon(t,\cdot)) + \epsilon j^1_\epsilon(q_\epsilon(t,\cdot)),$$

where $j^1_\epsilon$ is uniformly bounded, and $j^0(q) \geq 0$ with equality if and only if $q$ is a rest point of the limit dynamics. In this light, we only get Lyapunov stability *asymptotically* as $\epsilon \to 0$. The possible outcomes of this approach are still to be studied rigorously.

## 12.6 The Hamilton-Jacobi equation

In our context, the Hamilton-Jacobi approach has been introduced in [67] and then developed in [188, 154] to study the concentration effect for phenotypically structured PDEs. This consists in determining the possible Dirac distributions through the zeros of $u_\epsilon$ defined from the Hopf-Cole transform

$$u_\epsilon(t,x) = \epsilon \ln n_\epsilon(t,x).$$

In the mentioned works, the convergence of $u_\epsilon$ as $\epsilon$ goes to 0 is rigorously established and the limit $u$ satisfies a constrained Hamilton-Jacobi equation, using the theory of viscosity solutions (see [25] for an introduction). The constraint on the solution $u$ reads

$$\max_{x\in\mathbb{R}} u(t,x) = 0, \quad \forall t > 0$$

and comes from the control in $L^1$ of the total population. Then, some properties on the concentration points can be derived from the study of the constrained Hamilton-Jacobi equation and the solution $u$. In some particular cases, it is proved that the population density remains monomorphic, that is composed of a single Dirac mass, and then a form of canonical equation is derived, giving the dynamics of the dominant trait.

In the present work, a Hamilton-Jacobi structure arises in the different situations presented above. We show in this section different results obtained by applying the Hamilton-Jacobi approach, especially on the regularity of $u_\epsilon$. The main difficulties that we encounter are the time-dependency of the coefficients and their lack of regularity.

**Asymmetric fecundity:** we use the particular form

$$K_\epsilon(x,y,z) = B(y)\frac{1}{\epsilon}\alpha\left(\frac{x-z}{\epsilon}, y\right) \text{ with } \int \alpha(z',y)dz' = 1 \text{ for all } y,$$

and define

$$b_\epsilon(t,x) := R(x,\rho_\epsilon(t)), \quad q_\epsilon(t,y) = \frac{n_\epsilon(t,y)}{\rho_\epsilon(t)}. \tag{12.34}$$

In this case, the equation on $u_\epsilon$ reads

$$\partial_t u_\epsilon(t,x) = \int B(y)q_\epsilon(t,y)\int \alpha(z,y)e^{\frac{u_\epsilon(t,x-\epsilon z)-u_\epsilon(t,x)}{\epsilon}}dzdy - b_\epsilon(t,x), \tag{12.35}$$

and we compute the formal limiting equation

$$\begin{aligned}\partial_t u(t,x) &= \int B(y)q(t,y)\int \alpha(z,y)e^{-\partial_x u(t,x)\cdot z}dzdy - b(t,x)\\ &= \int B(y)q(t,y)\mathcal{L}[\alpha(\cdot,y)](\partial_x u(t,x))dy - b(t,x),\end{aligned} \tag{12.36}$$

with $\mathcal{L}[\alpha(\cdot\,,y)]$ the Laplace transform of $\alpha(\cdot\,,y)$ for all $y$:

$$\mathcal{L}[\alpha](p) := \int \alpha(z) e^{-p\,\cdot\,z} dz,$$

for $\alpha$ a probability density function.

**Asymmetric trait heredity:** The interest of this problem comes from the time- and trait-dependent coefficients of the Hamiltonian. We use the generic form

$$K_\epsilon(x,y,z) = G_\epsilon(x-z)K_1(x,y).$$

After the change of variable $z' = \frac{x-z}{\epsilon}$, the equation on $u_\epsilon$ reads

$$\partial_t u_\epsilon(t,x) = \frac{1}{\rho_\epsilon(t)} \int K_1(x,y) n_\epsilon(t,y) dy \cdot \int G(z') e^{\frac{u_\epsilon(t,x-\epsilon z')-u_\epsilon(t,x)}{\epsilon}} dz' - b_\epsilon(t,x). \tag{12.37}$$

For clarity, we define

$$a_\epsilon(t,x) := \int K_1(x,y) q_\epsilon(t,y) dy \geq 0. \tag{12.38}$$

At the limit $\epsilon \to 0$, we obtain the formal limiting equation

$$\begin{aligned} \partial_t u(t,x) &= a(t,x) \int G(z) e^{-\partial_x u(t,x)\,\cdot\,z} dz - b(t,x) \\ &= a(t,x) \mathcal{L}[G](\partial_x u(t,x)) - b(t,x), \end{aligned} \tag{12.39}$$

with $a$ and $b$ the formal limits of $a_\epsilon$ and $b_\epsilon$ defined in (12.38) and (12.34), and $\mathcal{L}[G]$ the Laplace transform of $G$. From now on, we choose $G$ such that its Laplace transform is well defined on $\mathbb{R}$.

In the case $G$ is the gaussian density, the equation on $u_\epsilon$ reads

$$\partial_t u_\epsilon(t,x) = a_\epsilon(t,x) \int \frac{1}{\sqrt{2\pi}} e^{-\frac{|z|^2}{2}} e^{\frac{u_\epsilon(t,x-\epsilon z)-u_\epsilon(t,x)}{\epsilon}} dz - b_\epsilon(t,x). \tag{12.40}$$

Then, passing formally to the limit $\epsilon \to 0$, we arrive at

$$\begin{aligned} \partial_t u(t,x) &= a(t,x) \int \frac{1}{\sqrt{2\pi}} e^{-\frac{|z|^2}{2}} e^{-\partial_x u(t,x)\,\cdot\,z} dz - b(t,x) \\ &= a(t,x) e^{(\partial_x u(t,x))^2/2} - b(t,x). \end{aligned}$$

The purpose of this section is to prove for these models the convergence of $u_\epsilon$ as $\epsilon$ vanishes. To this end, we derive a priori estimates on the sequence $u_\epsilon$ in order to use compactness arguments. The uniqueness of the solution to the limit equation has not been determined, thus we only derive convergence up to extraction of subsequences. Moreover, the stability result is not complete, since we do not have convergence results on the coefficients.

We mostly focus on the equation on $u_\epsilon$ (12.40) for the proof of Theorem 12.4, but the arguments are identical for the generic ATH case. The proof of the theorem in the AF case is similar and we also give the formal ideas where it is necessary.

**Assumptions:** We assume on the function $R$

$$\exists C_0 > 0,\ \forall \rho_m \leq \rho \leq \rho_M,\ \forall x \in \mathbb{R}, \quad R(x,\rho) \leq C_0(1+|x|), \tag{12.41}$$

$$\exists L_b > 0, \forall \rho_m \leq \rho \leq \rho_M,\ \forall x \in \mathbb{R}, \quad |\partial_x R(x,\rho)| \leq L_b. \tag{12.42}$$

We choose the positive function $K_1$ bounded

$$\exists \bar{K} > 0, \forall x,y \in \mathbb{R}, \quad K_1(x,y) \leq \bar{K}, \tag{12.43}$$

and such that,

$$\exists \lambda > 0, \exists C_\lambda > 0, \forall \epsilon > 0, t \geq 0, x \in \mathbb{R}, \quad e^{\frac{|\partial_x a_\epsilon(t,x)|}{\lambda a_\epsilon(t,x)}} \lambda a_\epsilon(t,x) \leq C_\lambda. \tag{12.44}$$

This assumption is satisfied for example when $K_1$ is bounded and there exists a constant $L_K$ such that

$$|\partial_x K_1(x,y)| \leq L_K |K_1(x,y)|, \quad \forall x,y \in \mathbb{R},$$

or, when $K_1$ induces a gaussian type distribution for $a_\epsilon$, that is,

$$a_\epsilon(t,x) \sim Ce^{\frac{-(x-m)^2}{\sigma^2}}.$$

We also assume on the initial condition

$$u_\epsilon^0(x) \leq -A|x| + C, \quad \|\partial_x u_\epsilon^0\| \leq L_0. \tag{12.45}$$

For the model with asymmetric fecundity, we assume that $B$ and $\alpha$ are positive and bounded. For both models under investigation, we prove Theorem 12.4.

### 12.6.1 A priori bounds

We begin with the estimates for the ATH case, and especially with a gaussian trait female heredity distribution.

**Lemma 12.4.** *Let $u_\epsilon$ be solution to equation (12.40). Then, there exist constants $C_1 > 0$ and $C_2 > 0$, such that for all $t > 0, x \in \mathbb{R}$ and $\epsilon > 0$ we have*

$$-C_1(1+t)(1+|x|) \leq u_\epsilon(t,x) \leq -A|x| + C_2(1+t).$$

We prove this lemma in the case of a gaussian trait female heredity distribution, but the argument exactly applies to equation (12.37) in the generic ATH case.

*Proof.* We first prove the lower bound

$$u_\epsilon(t,x) \geq -C_1(1+t)(1+|x|).$$

Indeed, because $a_\epsilon \geq 0$ and $\mathcal{L}(G) \geq 0$, we deduce from (12.41) that

$$\partial_t u_\epsilon \geq -b_\epsilon(t,x) \geq -C_0(1+|x|).$$

From (12.45) we obtain

$$u_\epsilon(t,x) \geq \inf_\epsilon u_\epsilon^0(0) - \inf_\epsilon \|\partial_x u_\epsilon^0\| - C_0 t(1+|x|).$$

Hence the lower bound.

We also derive the inequality

$$u_\epsilon(t,x) \leq -A|x| + C_2(1+t),$$

where $C_2 = \bar{K}\frac{1}{\sqrt{2\pi}}\int e^{-|z|^2/2}e^{A|z|}dz$. Indeed, defining $v(t,x) := -A|x| + C_2(1+t)$, we compute

$$\partial_t v(t,x) - a_\epsilon(t,x)\int \frac{1}{\sqrt{2\pi}}e^{-|z|^2/2}e^{\frac{v(t,x-\epsilon z)-v(t,x)}{\epsilon}}dz \geq C_2 - \bar{K}\frac{1}{\sqrt{2\pi}}\int e^{-|z|^2/2}e^{A|z|}dz \geq 0.$$

Thus, $v$ is a super-solution of (12.40), and since $u^0(x) \leq v(0,x)$ we deduce that $u_\epsilon(t,x) \leq v(t,x)$ by a comparison principle argument. □

We obtain the same kind of bounds for the asymmetric fecundity case.

**Lemma 12.5.** *Let $u_\epsilon$ be solution to equation (12.35). Then, there exist constants $C_1 > 0$ and $C_2 > 0$, such that for all $t > 0, x \in \mathbb{R}$ and $\epsilon > 0$ we have*

$$-C_1(1+t)(1+|x|) \leq u_\epsilon(t,x) \leq -A|x| + C_2(1+t),$$

*where $C_2 = \sup_y B(y)\int \alpha(z,y)e^{|A|z}dz$.*

### 12.6.2 Regularity in space

We prove the following

**Lemma 12.6.** *Let $u_\epsilon$ be the solution to the equation (12.40). For $\lambda > 0$ given by (12.44) and for all $t > 0, x \in \mathbb{R}$, we have*

$$|\partial_x u_\epsilon(t,x)| \leq \|\partial_x u_\epsilon^0\|_{L^\infty} + (C_\lambda + L_b)t + \lambda\left(\sup_\epsilon \|u_\epsilon^0\|_{L^\infty} + C_1(1+t)(1+|x|)\right).$$

*This implies that $u_\epsilon$ is Lipschitz in space, uniformly in $\epsilon$ and locally in time.*

*Proof.* We use the notations

$$p_\epsilon(t,x) = \partial_x u_\epsilon(t,x), \quad p(t,x) = \partial_x u(t,x).$$

Differentiating (12.40), $p_\epsilon$ satisfies

$$\begin{aligned}\partial_t p_\epsilon(t,x) = \partial_x a_\epsilon(t,x) \cdot \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{\frac{u_\epsilon(t,x-\epsilon z)-u_\epsilon(t,x)}{\epsilon}} dz \\ + a_\epsilon(t,x) \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{\frac{u_\epsilon(t,x-\epsilon z)-u_\epsilon(t,x)}{\epsilon}} \left(\frac{p_\epsilon(t,x-\epsilon z) - p_\epsilon(t,x)}{\epsilon}\right) dz - \partial_x b_\epsilon(t,x).\end{aligned}$$

Let $\lambda > 0$. We define

$$w_\epsilon^\lambda(t,x) = p_\epsilon(t,x) + \lambda u_\epsilon(t,x), \quad D_\epsilon(t,x,z) = \frac{u_\epsilon(t,x-\epsilon z) - u_\epsilon(t,x)}{\epsilon}.$$

Then, $w_\epsilon^\lambda$ satisfies

$$\begin{aligned}\partial_t w_\epsilon^\lambda = a_\epsilon \cdot \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{D_\epsilon(t,x,z)} \left(\frac{w_\epsilon^\lambda(t,x-\epsilon z) - w_\epsilon^\lambda(t,x)}{\epsilon}\right) dz \\ - \lambda\left[a_\epsilon \cdot \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{D_\epsilon(t,x,z)} (D_\epsilon(t,x,z) - 1)\right] dy \\ + \partial_x a_\epsilon \cdot \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{D_\epsilon(t,x,z)} dz - (\partial_x b_\epsilon + \lambda b_\epsilon).\end{aligned}$$

Then, using (12.42), we have

$$\begin{aligned}\partial_t w_\epsilon^\lambda - L_b - a_\epsilon \cdot \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{D_\epsilon} \left(\frac{w_\epsilon^\lambda(t,x-\epsilon z) - w_\epsilon^\lambda(t,x)}{\epsilon}\right) dz \\ \leq \int \frac{1}{\sqrt{\pi}} e^{-|z|^2} e^{D_\epsilon} \left[\partial_x a_\epsilon + \lambda a_\epsilon - \lambda a_\epsilon D_\epsilon\right] dz.\end{aligned}$$

Defining $f(D) := e^D(\partial_x a_\epsilon + \lambda a_\epsilon - \lambda a_\epsilon D)$, the maximum of $f$ on $\mathbb{R}$ is reached at $D^* := \frac{\partial_x a_\epsilon}{\lambda a_\epsilon}$ and equals

$$e^{\frac{\partial_x a_\epsilon}{\lambda a_\epsilon}} \lambda a_\epsilon \leq C_\lambda,$$

from (12.44). Then we have the upper bound

$$w_\epsilon^\lambda(t,x) \leq \max_{\mathbb{R}} w_\epsilon^\lambda(0,x) + Ct, \quad C = C_\lambda + L_b,$$

which implies the upper bound on $p_\epsilon$

$$p_\epsilon(t,x) \leq \|\partial_x u_\epsilon^0\|_{L^\infty} + Ct + \lambda\left(\sup_\epsilon \|u_\epsilon^0\|_{L^\infty} + C_1(1+t)(1+|x|)\right).$$

We have the same estimate for $-p_\epsilon$. □

For the AF model, we have the following estimate on the derivative in space of $u_\epsilon$:

**Lemma 12.7.** *Let $u_\epsilon$ be the solution of equation (12.35). Then, for all $t > 0, x \in \mathbb{R}$ and $\epsilon > 0$, we have*

$$|\partial_x u_\epsilon(t,x)| \leq \|\partial_x u_\epsilon^0\|_{L^\infty} + L_b t.$$

*This implies that $u_\epsilon$ is Lipschitz in space, uniformly in $\epsilon$ and locally in time.*

We address the limit equation

$$\partial_t p(t,x) = (-\partial_x p(t,x)) \int B(y) q(t,y) \int z\alpha(z,y) e^{-p(t,x)\cdot z} dz dy - \partial_x b(t,x), \tag{12.46}$$

and give formal arguments, since the proof for the $\epsilon$-level problem is similar to the one of the ATH case. We compute that $w(t) := \|\partial_x u_\epsilon^0\|_{L^\infty} + L_b t$ is a super-solution of (12.46). Since $p(0,x) \le w(0)$ for all $x \in \mathbb{R}$, we deduce that, from the comparison principle, $u_\epsilon$ is Lipschitz in space, uniformly in $\epsilon$ and locally in time.

### 12.6.3 Regularity in time

In the ATH case, since we proved that $u_\epsilon$ is uniformly Lipschitz in space locally in time, we can deduce that $\partial_t u_\epsilon$ is locally uniformly bounded.

**Lemma 12.8.** *Let $u_\epsilon$ be the solution to equation (12.37) and let $T > 0$ and $r > 0$ be fixed. Assume (12.42) and (12.43). Then, there exists $C(T,r) > 0$ such that, for all $t \in [0,T], x \in B(0,r)$, we have*

$$|\partial_t u_\epsilon| \le C(T,r) + \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))}.$$

*This implies that $u_\epsilon$ is Lipschitz in time, uniformly in $\epsilon$.*

*Proof.* Let $T > 0$ and $R > r > 0$ be fixed with R large enough. We choose some constants $L_1$ and $L_2$ such that

$$u_\epsilon(t,x) < -L_1, \quad \forall (t,x) \in [0,T] \times \mathbb{R}\backslash B(0,R),$$
$$|p_\epsilon| < L_2, \quad \forall (t,x) \in [0,T] \times B(0,R).$$

Then, we obtain for $t \in [0,T], x \in B(0,r)$,

$$\begin{aligned}|\partial_t u_\epsilon| \le & \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))} \\ & + \frac{1}{\rho_\epsilon(t)} \int K(x,z) n_\epsilon(t,z) dz \cdot \left( \int_{|x-\epsilon y|<R} e^{-|y|^2} e^{L_2 y} dy + \int_{|x-\epsilon y|>R} e^{-|y|^2} e^{\frac{u_\epsilon(t,x-\epsilon y)-u_\epsilon(t,x)}{\epsilon}} dy \right).\end{aligned}$$

Thus, for $\epsilon$ small enough, and assuming that

$$u_\epsilon(t,x) > -L_1, \quad \forall t \in [0,T], \forall x \in B(0,r),$$
$$u_\epsilon(t,x) < -L_1, \quad \forall t \in [0,T], \forall x \in \mathbb{R}\backslash B(0,R),$$

we have

$$\begin{aligned}|\partial_t u_\epsilon| &\le \overline{K} \left( \int_{|x-\epsilon y|<R} e^{-|y|^2} e^{L_2 y} dy + \int_{|x-\epsilon y|>R} e^{-|y|^2} e^{\frac{-L_1-u_\epsilon(t,x)}{\epsilon}} dy \right) + \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))} \\ &\le \overline{K} \left( \int e^{-|y|^2} e^{L_2 y} dy + \int_{|x-\epsilon y|>R} e^{-|y|^2} dy \right) + \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))} \\ &\le \overline{K} \left( \int e^{-|y|^2} e^{L_2 y} dy + \sqrt{\pi} \right) + \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))}.\end{aligned}$$

Hence the local uniform bound on $\partial_t u_\epsilon$. □

The proof is similar for the AF case.

**Lemma 12.9.** *Let $u_\epsilon$ be the solution to equation (12.35) and let $T > 0$ and $r > 0$ be fixed. Then, there exists $C(T,r) > 0$ such that, for all $t \in [0,T], x \in B(0,r)$, we have*

$$|\partial_t u_\epsilon| \le C(T,r) + \sup_{0\le\rho\le\rho_M} \|R(\,\cdot\,,\rho)\|_{L^\infty(B(0,r))}.$$

*This implies that $u_\epsilon$ is Lipschitz in time, uniformly in $\epsilon$.*

### 12.6.4 A more precise upper bound

The following argument concerns both cases and gives a sharper upper bound on $u_\epsilon$.

**Lemma 12.10.** *Let $u_\epsilon$ be the solution to equation (12.35) or (12.37). Then, for all $x, y \in \mathbb{R}$, we have*

$$u_\epsilon(t,x) \le \epsilon \ln\big(\rho_M m_{x,\frac{C(1+t)}{\epsilon}}\big),$$

*where $m_{x,A} > 0$ is the minimum on $\mathbb{R}$ of $g_{x,A} : y \mapsto A\frac{1+\max(|x|,|y|)}{1-e^{-|y-x|A(1+\max(|x|,|y|))}}$.*

*In addition, if $A > 0$ we have $A < m_{x,A} \le A + 3/2$. Thus, we obtain the global upper bound*

$$u_\epsilon(t,x) \le \epsilon \ln\big(\rho_M(3/2 + C(1+t)/\epsilon)\big) \xrightarrow{\epsilon\to 0} 0.$$

*Proof.* For all $z \in (x,y)$, by the mean value theorem there exists $\theta_\epsilon(t,x,z)$ between $x$ and $y$ such that

$$u_\epsilon(t,z) = u_\epsilon(t,x) + (z-x)\partial_x u_\epsilon(t,\theta_\epsilon(t,x,z)).$$

In addition, by the previous point there exists $C$ (independent of $t, x$ and $\epsilon$) such that for all $t, x$, $|\partial_x u_\epsilon(t,x)| \le C(1+t)(1+|x|)$. Hence

$$u_\epsilon(t,z) \ge u_\epsilon(t,x) - (z-x)C(1+t)\big(1+\max(|x|,|y|)\big).$$

Since we have, for $x < y$,

$$\int_x^y e^{\frac{u_\epsilon(t,z)}{\epsilon}} dz \le \rho_M,$$

we deduce that

$$\epsilon e^{\frac{u_\epsilon(t,x)}{\epsilon}} \frac{1 - e^{-(y-x)\frac{C(1+t)(1+\max(|x|,|y|))}{\epsilon}}}{C(1+t)\big(1+\max(|x|,|y|)\big)} \le \rho_M, \quad \forall y.$$

Then, we compute

$$u_\epsilon(t,x) \le \epsilon \ln\Big(\frac{\rho_M C(1+t)\big(1+\max(|x|,|y|)\big)}{\epsilon\big(1-e^{-(y-x)\frac{C(1+t)(1+\max(|x|,|y|))}{\epsilon}}\big)}\Big),$$

and this holds for all $y > x$. We can also choose $y < x$ and get in more generality

$$u_\epsilon(t,x) \le \epsilon \ln\Big(\frac{\rho_M C(1+t)\big(1+\max(|x|,|y|)\big)}{\epsilon\big(1-e^{-|y-x|\frac{C(1+t)(1+\max(|x|,|y|))}{\epsilon}}\big)}\Big) = \epsilon \ln\big(\rho_M g_{x,\frac{C(1+t)}{\epsilon}}(y)\big).$$

Observe that $g_{x,A}$ is positive and goes to $+\infty$ at $y = \pm\infty$ and at $y = x$. Minimizing in $y$, we find that

$$u_\epsilon(t,x) \le \epsilon \ln\big(\rho_M m_{x,\frac{C(1+t)}{\epsilon}}\big).$$

To conclude we first remark that if $A > 0$ and $x, y \in \mathbb{R}$, then we have

$$\frac{1+\max(|x|,|y|)}{1-e^{-|y-x|A(1+\max(|x|,|y|)}} > 1,$$

so $g_{x,A}(y) > A$ for all $y \in \mathbb{R}$ and thus $m_{x,A} > A$. Then, with $A > 0$ we also have

$$g_{1/A,A}(-1/A) = \frac{A+1}{1-e^{-2(1+A)}} \le A + 3/2,$$

which implies $m_{x,A} \le A + 3/2$. Thus, we obtain the global upper bound

$$u_\epsilon(t,x) \le \epsilon \ln\big(\rho_M(3/2 + C(1+t)/\epsilon)\big) \xrightarrow{\epsilon\to 0} 0.$$

□

The proof of Theorem 12.4 is achieved.

### 12.6.5 Discussion on the formal limiting equation

By Lipschitz regularity, as is classically proved with the Hamilton-Jacobi approach to adaptive dynamics (see [214]), the limit function $u$ satisfies the constraint

$$\max_{x\in\mathbb{R}} u(t,x) = 0, \quad \forall t > 0.$$

Then, when $u$ is differentiable at maximum points, we deduce that $\partial_t u$ equals 0 and, going back to (12.36) and (12.39), we obtain

$$\text{supp } \bar{n} \subset \{(t,x) \in (0,\infty)\times\mathbb{R} | B(x) - b(t,x) = 0\}, \quad \text{in case (AF)},$$

$$\text{supp } \bar{n} \subset \{(t,x) \in (0,\infty)\times\mathbb{R} | a(t,x) - b(t,x) = 0\}, \quad \text{in case (ATH)}.$$

It would be then interesting to determine the conditions required to have these null sets reduced to an isolated point. If, for all $t > 0$, we identify a unique point $\bar{x}(t)$ satisfying

$$B(\bar{x}(t)) - b(t,\bar{x}(t)) = B(\bar{x}(t)) - R(\bar{x}(t),\bar{\rho}(t)) = 0, \quad \text{in case (AF)},$$

$$a(t,\bar{x}(t)) - R(\bar{x}(t),\bar{\rho}(t)) = 0, \quad \text{in case (ATH)},$$

then the population is monomorphic, that composed of a single Dirac mass located on $\bar{x}(t)$. Provided some regularity properties on $u_\epsilon$, we could derive a canonical equation for both (AF) and (ATH).

Back to (nM), we define $\overline{n} \in \mathcal{M}_+(\mathbb{R})$ as an Evolutionary Stable Distribution (ESD) in the sense of [64, 123] by

$$K_0 * \overline{n} = \nu\overline{\rho}^2 \text{ on } supp(\overline{n}), \tag{12.47}$$

$$K_0 * \overline{n} \leq \nu\overline{\rho}^2 \text{ on } \mathbb{R}, \tag{12.48}$$

where $\overline{\rho} = \int \overline{n}$. The interest of the ESD concept is huge: it is readily established that a stationary solution to (nM) is asymptotically stable if and only if it satisfies (12.47) and (12.48).

If we assume that $K_0$ is radial-decreasing, then we prove that extreme points in $supp(\overline{n})$ (if it is bounded) cannot support a positive Dirac mass, by using (12.48). In particular, among all combinations of Dirac masses, only the single-point measure $\overline{n}_{\overline{x}}(x) := K_0(0)/\nu\delta_{x=\overline{x}}$ is an ESD. Indeed, assume that $\bar{n}$ is composed of $k \geq 2$ Dirac masses located on $(x_i)_{1\leq i\leq k}$, then defining

$$\overline{K}(x) := K_0 * \bar{n}(x) = \sum_{i=1}^{k} \rho_i K_0(x - x_i),$$

we deduce from (12.47) and (12.48) that $\overline{K}$ is maximal on the support of $\bar{n}$, that is the points $x_i$. With no loss of generality, we assume that the sequence $(x_i)$ is ordered and $x_1 = \min_i x_i$. Then, differentiating $\overline{K}$, we obtain

$$\overline{K}'(x_1) = \sum_{i\geq 1} \rho_i K_0'(x_1 - x_i) > 0,$$

which contradicts the optimality of $\overline{K}$ on the support of $\bar{n}$.

## 12.7 Conclusion and perspectives

We investigated adaptive dynamics for population dynamics model including sexual reproduction, when the trait is mainly inherited from the mother. We determined non-extinction conditions and a control on the total population. In the particular case of a saturation term $R$ depending only on the competition, we derived BV estimates on the total population. In general, estimating the variations of $\rho_\epsilon$ when $R$ depends on both trait variable and competition seems difficult, although an approach using a Lyapunov functional yields interesting results in some cases. More appropriate assumptions need to be considered.

Concerning the sequences $u_\epsilon = \epsilon \ln n_\epsilon$ associated to each model, we obtained local Lipschitz estimates uniform in $\epsilon$. To deduce the convergence of $u_\epsilon$ to the solution of the limiting Hamilton-Jacobi equation with constraint, we still need time compactness on the coefficients of (12.35) and (12.37). As a special case of both, for the Hamilton-Jacobi equation associated to the model

without mutations (gnM), if we provide some convergence result on $\int K(x,y) * n_\epsilon(t,y)/\rho_\epsilon(t)$ and on $\rho_\epsilon$, then, up to extraction of a subsequence, the limit function $u$ has an explicit formulation and its maximum points can be described. In general, Hamilton-Jacobi equations with time- and space-dependent coefficients are difficult to deal with when there is a lack of regularity. The authors in [150] developed a theory of stochastic viscosity solutions to tackle nonlinear stochastic PDEs. In particular, they prove existence, regularity and uniqueness results for the viscosity solution when the time-dependent coefficient of the Hamiltonian can be written as the derivative of a trajectory. This theory does not apply to our models since the coefficients in front of the gradient-dependent term are not under the form of a time derivative.

Another question is the determination of a convenient framework to observe Dirac concentrations. The convergence of the population distribution to a sum of Dirac masses illustrates the selection of well-adapted or dominant phenotypical traits. In [154], the Hamilton-Jacobi approach enables to characterize the dynamics of the dominant traits under specific assumptions of regularity. The required hypotheses to observe Dirac concentrations are to be clarified.

Another viewpoint has been recently developed in [194] using the Wasserstein distance to study a spatial infinitesimal model. It is proved that the sexual reproduction operator in the infinitesimal model induces a contraction for the Wasserstein distance on the phenotypical trait space, which enables to derive a macroscopic limit for the model, using also some parabolic estimates for the space regularity of the solution. It could be interesting to explore the Wasserstein approach to investigate general sexual population models, although at first sight the key inequality in [194] seems lost when the normalized gaussian kernel is replaced by a general $K(x,y,z)$ such that the total progeny $\int K(x,y,z)dx$ effectively depends on the traits of the parents, $y$ and/or $z$.

# Chapter 13

# Mathematical perspectives

Ignoranti quem portum petat nullus suus ventus est.

Seneca, *Epistulae morales ad Lucilium.*

In this chapter we collect four problems that we came across during the thesis but did not have time to tackle fully. We take this list of open problems as an opportunity for motivating further research in these topics.

## 13.1 Bubbles for elliptic systems

Motivated by the study of the two-dimensional competitive reaction-diffusion system from Section 4.3.3, we conjecture that the sharp threshold principle for scalar bistable equation extends to such bistable systems. The systems we consider are of the form

$$\begin{cases} \partial_t n_1 - D_1 \Delta n_1 = f_1(n_1, n_2) \text{ in } \mathbb{R}_+ \times \mathbb{R}^d, \\ \partial_t n_2 - D_2 \Delta n_2 = f_2(n_1, n_2) \text{ in } \mathbb{R}_+ \times \mathbb{R}^d, \\ (n_1(0,\cdot), n_2(0,\cdot)) = \mathbf{n}^0 \text{ in } \mathbb{R}^d, \end{cases} \tag{13.1}$$

where the non-linearities are such that the system is monotone with respect to some orthant $\mathcal{K}^0$. We further assume that there are exactly two stable steady states $\mathbf{0} \ll_{\mathcal{K}^o} \mathbf{E}_+$ for (13.1). We state a sharp threshold conjecture in this context:

**Conjecture 13.1.** *Let $(\mathbf{n}^0_\lambda)_{\lambda \geq 0}$ be an increasing (with respect to $\mathcal{K}^o$) family of initial data for the system (13.1), continuous in $L^1$ norm and such that $\mathbf{n}^0_0$ leads to $\mathbf{0}$. Then, there exists a unique $\lambda_0 \in (0, +\infty]$ such that:*

- *if $\lambda < \lambda_0$ then the solution converges to $\mathbf{0}$,*
- *if $\lambda > \lambda_0$ then the solution converges to $\mathbf{E}_+$.*

A first step towards the proof of such a conjecture could be the construction of "bubbles" or "propagules". In other words, we are looking for solutions to the Dirichlet elliptic problem in dimension $d = 1$, with $L > 0$:

$$\begin{cases} -D_1 n_1'' = f_1(n_1, n_2), \\ -D_2 n_2'' = f_2(n_1, n_2), \\ (n_1, n_2)(\pm L) = \mathbf{0}, \quad n_1, n_2 \in (\mathbf{0}, \mathbf{E}_+) \text{ in } (-L, L), \end{cases} \tag{13.2}$$

or to

$$\begin{cases} -D_1 n_1'' = f_1(n_1, n_2), \\ -D_2 n_2'' = f_2(n_1, n_2), \\ (n_1, n_2)(\pm L) = \mathbf{E}_+, \quad n_1, n_2 \in (\mathbf{0}, \mathbf{E}_+) \text{ in } (-L, L), \end{cases} \tag{13.3}$$

where by $n \in (\mathbf{0}, \mathbf{E}_+)$ we mean $\mathbf{0} \ll_{\mathcal{K}^o} n \ll_{\mathcal{K}^o} \mathbf{E}_+$.

Such problems are extensions to systems of the scalar ones solved in [182] and [183]. Existence, uniqueness and stability properties for monotone systems in dimension $N_d = 2$ have been obtained

under (too) specific conditions on the nonlinearities in [53] (see the introduction of the cited article for a historical and synthetic presentation of the techniques of proof). This result has been extended to cooperative systems in any dimension $N_d \geq 2$, under suitable assumptions in [234] (this approach has been pursued in [156]). However, to the best of our knowledge no result has been obtained yet for rather general nonlinearities such as these in (1.1).

First, we claim that in general there cannot be solutions to both problems (13.2) and (13.3) simultaneously. By classical results on competitive systems (see the discussion in Section 4.3.3), there exists a traveling wave $\psi := (\psi_1, \psi_2)$ connecting $\mathbf{0}$ at $-\infty$ to $\mathbf{E}_+$ at $+\infty$ and traveling at speed $c \in \mathbb{R}$. Using this particular solution and the comparison principle yields:

**Proposition 13.1.** *Assume $c \neq 0$. If there exists $L > 0$ such that (13.2) (resp. (13.3)) has a solution, then $c < 0$ (resp. $c > 0$) and for all $L > 0$, (13.3) (resp. (13.2)) has no solution.*

*In case $c < 0$ (resp. $c > 0$), if there exists $L_* \in (0, +\infty]$ such that for (13.2) (resp. (13.3)) has a solution then for all $L > L_*$ it also has a solution, denoted $\mathbf{n}_L$, such that if $L_* < L < L'$, $\mathbf{n}_L <_{\mathcal{K}^o} \mathbf{n}_{L'}$ (resp. $\mathbf{n}_L <_{\mathcal{K}^o} \mathbf{n}_{L'}$) on $(-L, L)$.*

*Proof.* Let $\underline{n} := (\underline{n}_1, \underline{n}_2)$ be a solution to (13.2) for some $L > 0$, which we extend by $\mathbf{0}$ on $\mathbb{R}\backslash(-L, L)$. Then $\underline{n}$ is a constant sub-solution to (13.1). By convergence to $\mathbf{E}_+$ at $+\infty$, there exists $\xi \in \mathbb{R}$ such that $\psi(\cdot - \xi) \geq_{\mathcal{K}^o} \underline{n}$. This inequality is preserved by the time-dynamics, and in particular $\mathbf{0}$ cannot be the invading state: if $c \neq 0$ then $c < 0$.

By the symmetric construction (extending by $\mathbf{E}_+$ on $\mathbb{R}\backslash(-L, L)$), we can show that any solution to (13.3) gives rise to a "bubble" super-solution, which prevents $\mathbf{E}_+$ from being the invading state and imposes $c > 0$ if $c \neq 0$.

In particular, assuming $c \neq 0$ implies that for any $L_1, L_2 > 0$, the existence of a solution to (13.2) with $L = L_1$ and to (13.3) with $L = L_2$ are incompatible, whence the first part of the result.

Then, the sub- and super- solution method exposed in Chapter 4, Proposition 4.5 for scalar elliptic equations extends to systems (by the same process of building monotone and bounded sequences of sub- and super-solutions). Assume that (13.2) has a solution $\mathbf{n}_L$ on $(-L, L)$. The constant function $\mathbf{E}_+ \mathbb{1}_{[-L', L']}$ is a super-solution for (13.2) posed on $(-L', L')$, and $\mathbf{n}_L$ extended by $\mathbf{0}$ on $(-L', -L) \cup (L, L')$ is a sub-solution. Since none of them are solutions, the iterative process converge to a solution which lies between $\mathbf{n}_L$ and $\mathbf{E}_+$ and can be equal to none of them. Because of this property, the set of $L > 0$ such that (13.2) has a solution is either empty or a half-line containing its infimum $[L_*, +\infty) \subset \mathbb{R}_+$.

The argument is symmetrical for (13.3), where naturally the solution on $(-L, L)$ is extended by $\mathbf{E}_+$ on $(-L', -L) \cup (L, L')$. □

**Remark 13.1.** *Note that the proof applies also in the scalar case. The bubbles studied in Chapter 7 (see in particular Theorem 7.1 and Section 7.3) also possess the comparison property. Under the additional assumptions of Proposition 7.8, we know that there is exactly one bubble of radius $L_* > 0$ and two bubbles of radius $L$ for all $L > L_*$. In this case, applying Proposition 13.1 we can extract one subfamily $(p_L)_{L > L_*}$ such that $p_L$ has radius $L$ and the $p_L$ are ordered: for $L < L'$, $p_L < p_{L'}$ on $(-L, L)$.*

By continuity of $c$ with respect to perturbations of $f_1$ and $f_2$ (or $D_1$ and $D_2$), we can indeed deduce from Proposition 13.1 that generically, (13.2) and (13.3) cannot be solved simultaneously.

As noted in Section 4.3.3, the sign of $c$ is usually not simple to determine. With the bubble viewpoint, it amounts to checking which Dirichlet problem has solutions among (13.2) and (13.3).

Numerical simulations lead us to the following conjecture, which may hold under specific additional assumptions on $f_1$ and $f_2$:

**Conjecture 13.2.** *Assume $c > 0$ (resp. $c < 0$). There exists $L_* > 0$ such that (13.3) (resp. (13.2)) has*

- *0 non-negative solution for $L < L_*$,*
- *1 non-negative solution for $L = L_*$,*
- *2 non-negative solutions for $L > L_*$.*

**Remark 13.2.** *Problems such as (13.2), (13.3) can also arise from small parameter reduction of a degenerate reaction-diffusion system featuring non-diffusing compartments (mixing for instance Chapters 8 and 5). Indeed, a simple mosquito population model with spatial structure on adults (a) and juveniles (j), infected (subscript i) or not (subscript u) by a CI-inducing Wolbachia could read*

$$\begin{cases} \partial_t j_u(t,x) = b^u a_u(t,x)\big(1 - s_h \frac{a_i(t,x)}{a_u(t,x)+a_i(t,x)}\big)\big(1 - \frac{J(t,x)}{K}\big) - (\nu_j^u + \mu_j^u) j_u(t,x), \\ \partial_t j_i(t,x) = b^i a_i(t,x)\big(1 - \frac{J(t,x)}{K}\big) - (\nu_j^i + \mu_j^i) j_i(t,x), \\ \partial_t a_u(t,x) - D\Delta a_u(t,x) = \nu_j^u j_u(t,x) - \mu_a^u a_u(t,x), \\ \partial_t a_i(t,x) - D\Delta a_i(t,x) = \nu_j^i j_i(t,x) - \mu_a^i a_i(t,x), \\ J(t,x) = j_u(t,x) + j_i(t,x). \end{cases} \tag{13.4}$$

*Under the scaling assumption $b^z = b_0^z/\epsilon$ ($z \in \{u,i\}$), when $\epsilon \to 0$ we can prove system reduction for (13.4), as is done in Chapter 5. The study of sub- or super-solutions to (13.4), even in the reduced system, naturally leads to problems such as (13.2), (13.3).*

## 13.2 Wave-delaying

In Chapter 6, when $C > 0$ and $L < L_*(C)$, we have proved that there is no $(C, L)$-barrier. Can we construct an entire solution in this case, converging to the bistable traveling wave at $t \to \pm\infty$? Can we compute the associated delay? These interesting and natural questions were raised by Luca Rossi after the communication of the results of Chapter 6.

Until now we have not obtained satisfactory estimations, but for the sake of completeness we detail below an approach which seems to indicate the possibility of a quantitative answer.

For simplicity we consider only the cubic bistable nonlinearity and study the following equation:

$$\partial_t p - \partial_{xx} p - \epsilon C \mathbb{1}_{[0,2L]}(x)\partial_x p = p(1-p)(p-\theta), \tag{13.5}$$

where $\epsilon, C, L > 0$ and $\theta \in (0,1)$.

The unique (up to translation) traveling wave solution (connecting 1 at $-\infty$ to 0 at $+\infty$) of the homogeneous problem (as $C = 0$ or $L = 0$) is given by the profile $U$ and speed $c$:

$$U(\xi) := \frac{1}{1 + e^{\xi/\sqrt{2}}}, \quad c = \frac{1-2\theta}{\sqrt{2}}. \tag{13.6}$$

Following the approach in [224] (initiated by Fife and McLeod in [85]), one possibility is to look first for a sub-solution to (13.5) of the form $\underline{u}(t,x) = U(x - ct + \xi_\epsilon(t)) - \epsilon v(t,x)$. The interest of such an ansatz, in the context of wave-delaying, is that $\xi_\epsilon(t)/c$ is an underestimation of the delay at time $t$. If we can find limits $\xi_\pm^\epsilon = \lim_{t\to\pm\infty} \xi_\epsilon(t)$, then $\overline{\xi} := (\xi_+^\epsilon - \xi_-^\epsilon)/c = \int_{-\infty}^{+\infty} \xi_\epsilon'(t)dt/c$ is an overestimation of the delay induced by the obstacle $\epsilon C \mathbb{1}_{[0,2L]}$.

By definition, $\underline{u}$ is a sub-solution if and only if

$$-\epsilon\big(\partial_t v - \partial_{xx} v - \epsilon C \mathbb{1}_{[0,2L]}\partial_x v\big) + \xi_\epsilon'(t)U' - \epsilon C U' \mathbb{1}_{[0,2L]} \le f(U - \epsilon v) - f(U),$$

where $f : p \mapsto p(1-p)(p-\theta)$.

Recalling that $U' < 0$, we get

$$\xi_\epsilon'(t) \ge \frac{f(U-\epsilon v) - f(U) + \epsilon\big(\partial_t v - \partial_{xx} v - \epsilon C \mathbb{1}_{[0,2L]}\partial_x v\big)}{U'} + \epsilon C \mathbb{1}_{[0,2L]}..$$

We introduce

$$\Phi_x^\epsilon(t,\xi) := \frac{f(U(x-ct+\xi) - \epsilon v(t,x)) - f(U(x-ct+\xi)) + \epsilon\big(\partial_t v(t,x) - \partial_{xx} v(t,x) - \epsilon C \mathbb{1}_{[0,2L]}(x)\partial_x v(t,x)\big)}{U'(x-ct+\xi)},$$

so that the previous inequality now reads

$$\forall x \in \mathbb{R}, \quad \xi_\epsilon'(t) \ge \Phi_x^\epsilon(t, \xi_\epsilon(t)) + \epsilon C \mathbb{1}_{[0,2L]}(x).$$

To obtain as small delay overestimation as possible, it makes sense to define

$$\xi_\epsilon'(t) = \sup_{x\in\mathbb{R}} \big(\Phi_x^\epsilon(t,\xi_\epsilon(t)) + \epsilon C \mathbb{1}_{[0,2L]}(x)\big).$$

At this stage, one still has the freedom of defining $v$.

It seems natural to take $v$ with support in $[0,2L]$, and when dealing only with the time-forward problem, exponentially decreasing over time, so we take $v(t,x) = e^{-\gamma t}w(x)$. In order that $\Phi_x^\epsilon(t,\xi) \le -\epsilon C$ as $t$ goes to $+\infty$ (so that the delay stabilizes), since $U(x-ct-\xi)$ converges uniformly to 1 as $t\to+\infty$ for $x\in[0,2L]$ and a bounded delay $\xi$ and since $U'(x-ct-\xi)$ decays as $e^{-ct/\sqrt{2}}$, it suffices to take $\gamma < c/\sqrt{2} = 1/2-\theta$, and for $w$ the solution to

$$\begin{cases} -\partial_{xx}w - \epsilon C\partial_x w - (f'(1)+\gamma)w = 1 \text{ in } (0,2L), \\ w(0) = w(2L) = 0, \quad w > 0 \text{ in } (0,2L), \end{cases} \tag{13.7}$$

where for our particular choice we have $f'(1) = -(1-\theta)$ so that $-(f'(1)+\gamma) > 1/2$. We find

$$w(x) = 1 + e^{-\frac{\epsilon C}{2}x}\Big(\frac{\cosh(L\lambda) - e^{\epsilon CL}}{\sinh(L\lambda)}\sinh\big(\frac{\lambda x}{2}\big) - \cosh\big(\frac{\lambda x}{2}\big)\Big),$$

with $\lambda := \sqrt{4(1-\theta-\gamma)+\epsilon^2C^2}$, which satisfies $w>0$ on $(0,2L)$ at least for $\epsilon>0$ small enough.

With this ansatz, we obtain

$$\Phi_x^\epsilon(t,\xi) = \epsilon e^{-\gamma t}\frac{1+w(x)\big(2\theta+3U^2(x-ct+\xi)-2(1+\theta)U(x-ct+\xi)-1\big)}{U'(x-ct+\xi)} + \epsilon^2\frac{v^2(t,x)(1+\theta-3U(x-ct+\xi))}{U'(x-ct+\xi)} + \epsilon^3\frac{v^3(t,x)}{U'(x-ct+\xi)}.$$

Keeping only the first-order term in $\epsilon$ yields coefficient

$$\Xi_\gamma(t,\xi) := e^{-\gamma t}\max_{x\in[0,2L]}\frac{1+w(x)\big(2\theta+3U^2(x-ct+\xi)-2(1+\theta)U(x-ct+\xi)-1\big)}{U'(x-ct+\xi)},$$

with

$$\xi_\epsilon'(t) = \epsilon\big(\Xi_\gamma(t,\xi_\epsilon(t)) + C\big)_+.$$

At this stage, one should try to estimate $\Xi_\gamma$ to go further. At least, we know that as $\epsilon$ goes to 0, $w$ converges to

$$w_0(x) := 1 + \frac{\cosh(2L\sqrt{1-\theta-\gamma})-1}{\sinh(2L\sqrt{1-\theta-\gamma})}\sinh(x\sqrt{1-\theta-\gamma}) - \cosh(x\sqrt{1-\theta-\gamma}),$$

and that $\Xi_\gamma(t,\xi) < 0$ as soon as $ct-\xi$ is large enough.

Let us fix $W = \max_{[0,2L]} w(x) > 0$. Then it is readily seen (since the sign of the terms are known) that as a rough estimate,

$$\Xi_\gamma(t,\xi) \le \sqrt{2}e^{-\gamma t}\max_{x\in[0,2L]}\frac{(1+V)^2 + W\big(3-2(1+\theta)(1+V)+(2\theta-1)(1+V)^2\big)}{-V},$$

with $V = e^{(x-ct-\xi)/\sqrt{2}}$. In the subcase $W(1-2\theta)\ge 1$, we can deduce that

$$\Xi_\gamma(t,\xi) \le \sqrt{2}e^{-\gamma t}\Big(\big(W(1-2\theta)-1\big)e^{(2L-ct+\xi)/\sqrt{2}} + 2W(2-\theta) - e^{(ct-\xi-2L)/\sqrt{2}}\Big).$$

Hence we get a super-solution for $\xi$ by solving an equation of the form $Y = e^{\xi/\sqrt{2}}$,

$$Y' = \epsilon\big(\alpha Y^2 e^{-\kappa_1 t} + \beta Y - \eta e^{\kappa_2 t}\big)_+, \quad Y(0) = Y_0 \in (1,+\infty), \tag{13.8}$$

where $\alpha,\beta,\kappa_1,\kappa_2 > 0$, with $\kappa_1 = \kappa_2 + 2\gamma$. This equation reaches a constant value at the first time $t\ge 0$ such that

$$Y(t) \le \frac{-\beta+\sqrt{\beta^2+4\alpha\eta e^{-2\gamma t}}}{2\alpha}e^{\kappa_1 t} \sim_{t\to+\infty} \frac{\eta}{\beta}e^{\kappa_2 t}.$$

In details, we have

$$\alpha = (W(1-2\theta)-1)e^{L\sqrt{2}},\ \beta = 2W(2-\theta)+\frac{C}{\sqrt{2}},\ \eta = e^{-L\sqrt{2}},\ \kappa_1 = \frac{c}{\sqrt{2}}+\gamma,\ \kappa_2 = \frac{c}{\sqrt{2}}-\gamma.$$

The solution is non-constant as soon as $Y_0 \geq \frac{-\beta+\sqrt{\beta^2+4\alpha\eta}}{2\alpha}$, that is if initially the sub-solution is rather on the left of the obstacle ($\xi_0$ large enough).

Assume this equation (13.8) yields some limit value $Y_\epsilon^\infty(Y_0)$ as $t \to +\infty$. In the limit $\epsilon \to 0$, the expected delay overestimation is of order 1 in $\epsilon$, with first-order coefficient equal to

$$\overline{\xi}_1(\xi_0) := \lim_{\epsilon\to 0} \sqrt{2}\log\left(e^{-\xi_0/\sqrt{2}}Y_\epsilon^\infty(e^{\xi_0/\sqrt{2}})\right)/\epsilon.$$

In order to go further in the resolution of (13.8), we introduce $Z = Ye^{-\kappa_1 t}$, so that the equation becomes, as long as the solution $Y$ is non-constant:

$$Z' = \epsilon\alpha Z^2 + (\epsilon\beta - \kappa_1)Z - \epsilon\eta e^{-2\gamma t}, \tag{13.9}$$

that is as long as

$$Z(t) > \frac{-\beta+\sqrt{\beta^2+4\alpha\eta e^{-2\gamma t}}}{2\alpha}.$$

It is well-known that if a particular solution $Z_1$ to (13.9) is given, then $Z = Z_1 + 1/v$ is also a solution, where $v$ solves the linear equation

$$v' = -(\epsilon\beta - \kappa_1 + 2\epsilon\alpha Z_1)v - \epsilon\alpha.$$

Then, we would like to approximate solutions to (13.9) as $\epsilon \to 0$. Using a formal approach with the Duhamel integral formula, we develop:

$$Z_\epsilon(t) = Z_0 e^{-\kappa_1 t} + \epsilon\Big(\beta t Z_0 e^{-\kappa_1 t} + \alpha\frac{1-e^{-\kappa_1 t}}{\kappa_1} - \eta\frac{e^{\kappa_2 t}-1}{\kappa_2}\Big) + o(\epsilon).$$

Then, we are looking for the first time $t = t(\epsilon) > 0$ such that

$$Z_0 e^{-\kappa_1 t} + \epsilon\Big(\beta t Z_0 e^{-\kappa_1 t} + \alpha\frac{1-e^{-\kappa_1 t}}{\kappa_1} - \eta\frac{e^{\kappa_2 t}-1}{\kappa_2}\Big) = \frac{-\beta+\sqrt{\beta^2+4\alpha\eta e^{-2\gamma t}}}{2\alpha}.$$

Assume that $t(\epsilon)$ has a limit $\bar{t}$ as $\epsilon \to 0$, namely the solution to

$$Z_0 e^{-\kappa_1 \bar{t}} = \frac{-\beta+\sqrt{\beta^2+4\alpha\eta e^{-2\gamma \bar{t}}}}{2\alpha}.$$

Then, with $Z_0 = e^{\xi_0/\sqrt{2}}$,

$$\begin{aligned}
\overline{\xi}_1(\sqrt{2}\log(Z_0)) &= \sqrt{2}\lim_{\epsilon\to 0}\frac{1}{\epsilon}\log\Big(\frac{Z_\epsilon(t(\epsilon))e^{\kappa_1 t(\epsilon)}}{Z_0}\Big),\\
&= \sqrt{2}\lim_{\epsilon\to 0}\frac{1}{\epsilon}\log\Big(1+\epsilon\Big(\beta t(\epsilon) + \frac{\alpha}{Z_0}\frac{e^{\kappa_1 t(\epsilon)}-1}{\kappa_1} - \frac{\eta}{Z_0}\frac{e^{(\kappa_1+\kappa_2)t(\epsilon)}-e^{\kappa_1 t(\epsilon)}}{\kappa_2}\Big)\Big),\\
&= \sqrt{2}\Big(\beta\bar{t} + \frac{\alpha}{Z_0}\frac{e^{\kappa_1\bar{t}}-1}{\kappa_1} - \frac{\eta}{Z_0}\frac{e^{(\kappa_1+\kappa_2)\bar{t}}-e^{\kappa_1\bar{t}}}{\kappa_2}\Big).
\end{aligned}$$

This yields an analytical overestimation of the delay (in forward time) $\xi_0 \to \overline{\xi}_1$ for any suitable $\gamma$.

Symmetrically, we can build a super-solution using the ansatz $\overline{u}(t,x) = U(x-ct+\zeta_\epsilon(t))+\epsilon v(t,x)$. Then, $\overline{u}$ is a super-solution if and only if

$$\epsilon\big(\partial_t v - \partial_{xx}v - \epsilon C\mathbb{1}_{[0,2L]}\partial_x v\big) + \zeta_\epsilon'(t)U' - \epsilon C\mathbb{1}_{[0,2L]}(x)U' \geq f(U+\epsilon v) - f(U),$$

or equivalently

$$\zeta_\epsilon'(t) \leq \epsilon C\mathbb{1}_{[0,2L]}(x) + \underbrace{\frac{f(U+\epsilon v)-f(U)-\epsilon\big(\partial_t v - \partial_{xx}v - \epsilon C\mathbb{1}_{[0,2L]}\partial_x v\big)}{U'}}_{=:\Psi_x^\epsilon(t,\zeta)}.$$

To obtain a delay underestimation as large as possible, it makes sense to define

$$\zeta_\epsilon'(t) = \inf_{x\in\mathbb{R}} \big(\Psi_x^\epsilon(t,\zeta_\epsilon(t)) + \epsilon C \mathbb{1}_{[0,2L]}(x)\big).$$

Using a similar ansatz as before for $v$: $v(t,x) = e^{-\gamma t}w(x)$ (with the same $w$), this time with $\gamma > c/\sqrt{2}$, we obtain

$$\Psi_x^\epsilon(t,\zeta) = \epsilon e^{-\gamma t}\frac{-1 + w(x)\big(1-2\theta+2(1+\theta)U(x-ct+\zeta) - 3U^2(x-ct+\zeta)\big)}{U'(x-ct+\zeta)} + \epsilon^2 e^{-2\gamma t}\frac{w^2(x)(1+\theta-3U(x-ct+\xi))}{U'(x-ct+\zeta)} - \epsilon^3 e^{-3\gamma t}\frac{w^3(x)}{U'(x-ct+\zeta)}.$$

Other choices are possible. For instance, choosing $\gamma < c/\sqrt{2}$ and $v(t,x) = e^{-\gamma t}z(x)$ where $z$ solves

$$\begin{cases} -\partial_{xx}z - \epsilon C\partial_x z - (f'(1)+\gamma)z = 0 \text{ in } (0,2L), \\ z(0) = z(2L) = 0, \quad z > 0 \text{ in } (0,2L), \end{cases}$$

would yield

$$\Psi_x^\epsilon(t,\zeta) = \epsilon e^{-\gamma t}\frac{z(x)\big(1-2\theta+2(1+\theta)U(x-ct+\zeta) - 3U^2(x-ct+\zeta)\big)}{U'(x-ct+\zeta)} + \epsilon^2 e^{-2\gamma t}\frac{z^2(x)(1+\theta-3U(x-ct+\xi))}{U'(x-ct+\zeta)} - \epsilon^3 e^{-3\gamma t}\frac{z^3(x)}{U'(x-ct+\zeta)}.$$

From this, similar computations as above could be performed to deduce delay underestimations.

We note that although we developed the computations only in the case of a cubic nonlinearity given by (13.6), this approach also applies to general bistable reaction terms, upon replacing the exponential decay rate of $U'(K-ct)$, $c/\sqrt{2}$, by the general formula $\frac{-c+\sqrt{c^2-4f'(1)}}{2}$. In this setting, when looking for a sub-solution with the same ansatz as in the cubic case, $v(t,x) = w(x)e^{-\gamma t}$ where

$$w \text{ solves (13.7)}, \quad 0 < \gamma < \frac{-c+\sqrt{c^2-4f'(1)}}{2},$$

we obtain that the first-order term (in $\epsilon$) of $\Phi_x^\epsilon(t,\xi)$ has coefficient

$$\Xi_\gamma(t,\xi) = e^{-\gamma t}\max_{x\in[0,2L]}\frac{1-w(x)\big(f'(U(x-ct+\xi)) - f'(1)\big)}{U'(x-ct+\xi)},$$

and it remains to solve

$$\xi_\epsilon'(t) = \epsilon\big(\Xi_\gamma(t,\xi_\epsilon(t)) + C\big)_+$$

to obtain a delay over-estimation. Explicitly,

$$w(x) = 1 + e^{-\frac{\epsilon C}{2}x}\Big(\frac{\cosh(L\lambda) - e^{\epsilon CL}}{\sinh(L\lambda)}\sinh\big(\frac{\lambda x}{2}\big) - \cosh\big(\frac{\lambda x}{2}\big)\Big),$$

with $\lambda := \sqrt{4(-f'(1)-\gamma) + \epsilon^2 C^2}$, which satisfies $w > 0$ on $(0,2L)$ at least for $\epsilon > 0$ small enough.

## 13.3 Time-scales and limits for controlled slow-fast dynamics

We recall the notations from Chapter 10, where we considered a singular limit for a control system modeling mosquito population replacement by releases of *Wolbachia*-carrying individuals. Let $n_1^\epsilon$ (resp. $n_2^\epsilon$) denote the wild (resp. introduced) population, with fecundity $b_1^0/\epsilon$ (resp. $b_2^0/\epsilon$) and death rate $d_1$ (resp. $d_2$), with $d_i, b_i^0 > 0$ ($i \in \{1,2\}$), in an environment with capacity $K > 0$, and uni-directional sterile crossings (due to cytoplasmic incompatibility) with rate $s_h \in (0,1]$. We recall the definition 10.8:

$$\mathcal{U}_{T,C,M} = \{u \in L^\infty([0,T]), \quad 0 \le u \le M \text{ a.e. }, \int_0^T u(t)\,dt \le C\}.$$

Then, $u \in \mathcal{U}_{T,C,M}$ models the flux of released mosquitoes. The two-dimensional system under study reads:

$$\begin{cases} \dfrac{dn_1^\epsilon}{dt} = \dfrac{b_1^0}{\epsilon} n_1^\epsilon (1 - s_h \dfrac{n_2}{n_1^\epsilon + n_2^\epsilon})(1 - \dfrac{n_1^\epsilon + n_2^\epsilon}{K}) - d_1 n_1^\epsilon, \quad n_1^\epsilon(0) = K(1 - \epsilon \dfrac{d_1}{b_1^0}), \\ \dfrac{dn_2^\epsilon}{dt} = \dfrac{b_2^0}{\epsilon} n_2^\epsilon (1 - \dfrac{n_1^\epsilon + n_2^\epsilon}{K}) - d_2 n_2^\epsilon + u, \quad n_2^\epsilon(0) = 0. \end{cases} \tag{13.10}$$

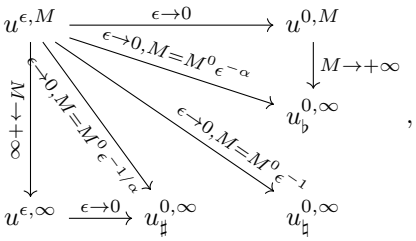
The non-linear functions in right-hand sides are denoted respectively $f_1^\epsilon$ and $f_2^\epsilon$. We define the set of reachable states:

$$\mathcal{X}_{T,C,M}^\epsilon := \{\big(n_1^\epsilon(t), n_2^\epsilon(t)\big) \in \mathbb{R}^2, \quad u \in \mathcal{U}_{T,C,M},\, t \in [0,T]\} \subset \mathbb{R}_+^2.$$

Given a smooth function $G^\epsilon : \mathbb{R}_+^2 \to \mathbb{R}$ such that almost-everywhere in $\mathcal{X}_{T,C,M}^\epsilon$, $\partial_1 G^\epsilon \geq 0$ and $\partial_2 G^\epsilon \leq 0$, we define the functional $J_{T,C,M}^\epsilon : u \mapsto G^\epsilon\big(n_1^\epsilon(T), n_2^\epsilon(T)\big)$. Our aim is to minimize $J_{T,C,M}^\epsilon$ on $\mathcal{U}_{T,C,M}$.

Expanding upon the conclusion of Chapter 10, we derive formally limit problems when $M$ scales in $M^0/\epsilon^\alpha$, for $\alpha > 0$. By taking this scaling, we actually compare the two short time-scales for the release process: *time concentration* of controls on the first hand and the *fast environment-filling* due to large fecundity on the other hand. In other words, we compare the short time-scale of a slow-fast system and the time-scale at which a control is applied on this system.

The following diagram is a comprehensive picture of the problems, for any $\alpha \in (0,1)$:




where $u^{\epsilon,M}$ is a minimizer of $J_{T,C,M}^\epsilon$ in $\mathcal{U}_{T,C,M}$. We expect that $u_\sharp^{0,\infty} \neq u_\flat^{0,\infty}$ in general.

We have already solved (essentially with Proposition 10.1) the situation $\alpha \in (0,1)$, when the environment-filling is much faster than the release process. In this case, the limit problem appears to have in general exactly one solution, given by a Dirac mass either at $t = 0$ or at $t = T$ (see Theorem 10.1, written for $G^\epsilon(x,y) = \frac{1}{2}x^2 + \frac{1}{2}(K(1 - \epsilon d_2/b_2^0) - y)_+^2$, which can be extended easily to the general final criterion under study).

However it would be necessary for applications to understand also the converse situation, where the control can be concentrated on a time-scale typically shorter than the free system's characteristic time-scale. When $\alpha \in (1, +\infty)$ (or $\alpha = 1$ and $M^0$ is large enough, but we do not explore this case in details here) we can use an interesting qualitative property of the minimizers, stated and proved in Section 10.B, Proposition 10.3.

We formulate an additional conjecture, inspired by numerical results and formal arguments (see Section 10.3.2):

**Conjecture 13.3.** *Let $u^*$ be a local minimizer of $J_{T,C,M}^\epsilon$ in $\mathcal{U}_{T,C,M}$. The set $I_* = \{u^* \in (0,M)\}$ is an interval.*

Combining Proposition 10.3 and Conjecture 13.3 reduces the problem to four dimensions: the bounds $t_0, t_1$ in $I_M = [0,t_0] \cup [t_1,T]$ and the interval $I_*$ which we denote $[t_2,t_3]$ characterize the minimizers.

So far, we have understood the arrows leading to $u_\flat^{0,\infty}$, by Proposition 10.1. We also have a good knowledge of $u^{\epsilon,\infty}$ as noted above, assuming Conjecture 13.3:

$$u^{\epsilon,\infty} = C_0^\epsilon \delta_0 + C_T^\epsilon \delta_T + \big(H^\epsilon(\mathbf{n}^\epsilon(t)) - f_2^\epsilon(\mathbf{n}^\epsilon(t))\big)\mathbb{1}_{[t_2(\epsilon),t_3(\epsilon)]}$$

for some $(C_0^\epsilon, C_T^\epsilon) \in \mathcal{T}_C := \{(C_1,C_2) \in [0,C]^2, \quad C_1 + C_2 \leq C\}$ and $0 \leq t_2(\epsilon) \leq t_3(\epsilon) \leq T$. Therefore along the horizontal arrow leading to $u_\sharp^{0,\infty}$ we can actually take controls in a set isometric to a compact subset of $\mathbb{R}^4$. We claim that the limit criterion can be expressed as $J_\sharp(C_0^0, C_T^0, t_H) =$

$(1-p(T))^2$, assuming that $t_2(\epsilon), t_3(\epsilon)$ converge to $t_H$ as $\epsilon \to 0$, upon defining carefully $p$. Morally, on most of $[0,T]$ the limit $p$ simply satisfies

$$\dot{p} = p(1-p)\frac{d_1 b_2^0 - d_2 b_1^0(1-s_h p)}{b_1^0(1-p)(1-s_h p) + b_2^0 p} =: f(p).$$

However, we need to understand precisely the limit (layer) problems as $\epsilon \to 0$, happening at $t=0$, $t=t_H$ and $t=T$, and possibly combining if $t_H = 0$ or $t_H = T$. To this aim, we construct below two mappings

$$\Phi_1 : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}^2, \quad \Phi_2 : \mathbb{R}^2 \to \mathbb{R},$$

where $\Phi_1$ represents the jump (as $\epsilon \to 0$) due to a singular arc with prescribed limit mass, and $\Phi_2$ represents the relaxation of the slow-fast system in $(n^\epsilon, p^\epsilon)$ towards its attracting curve (when $u^\epsilon$ vanishes) its value being the limit projection (as $\epsilon \to 0$) of $p$.

In details we let $H := \lim_{\epsilon\to 0} \epsilon H^\epsilon$ and

- $\Phi_1(C, x_0, y_0) = (x(t_f), y(t_f))$ where $(x,y)$ solves

$$\begin{cases} \dot{x} = b_1^0 x(1 - s_h \frac{y}{x+y})(1 - \frac{x+y}{K}), & x(0) = x_0, \\ \dot{y} = H(x,y), & y(0) = y_0, \end{cases}$$

  and $t_f$ is given by $\int_0^{t_f} \big(H(x(t),y(t)) - b_2^0 y(t)(1 - \frac{x(t)+y(t)}{K})\big)dt = C$;

- $\Phi_2(x_0, y_0) = z(\infty)$ where $(w,z)$ solves

$$\begin{cases} \dot{w} = -w(1-w)\big(b_1^0(1-z)(1-s_h z) + b_2^0 z\big), & w(0) = 1 - \frac{x_0+y_0}{K}, \\ \dot{z} = wz(1-z)\big(b_2^0 - b_1^0(1-s_h z)\big), & z(0) = \frac{y_0}{x_0+y_0}. \end{cases}$$

To find $p(0^+)$ we need to compute either $\Phi_2(K, C_0^0)$ (in case $u$ is a Dirac mass at 0 with limit mass $C_0^0$) or $\Phi_2 \circ \Phi_1(C_0^H, K, C_0^0)$ in case $u$ is a Dirac mass at 0 with limit mass $C_0^0$ immediately followed by a singular arc with limit mass $C_0^H$.

Similar computations give the jumps of $p$ at times $t_H$ and $T$ where the mass of $u^\epsilon$ concentrates as $\epsilon \to 0$, depending on the type of control: $u^\epsilon = M/\epsilon^\alpha$ (at $T$, relying on $\Phi_2$) or $u^\epsilon = H_\epsilon(\mathbf{n}^\epsilon)$ (at $t_H$ relying on both $\Phi_2$ and $\Phi_1$ since the system needs to be projected immediately after the control is stopped).

On the intervals $(0, t_H)$ and $(t_H, T)$, one simply needs to solve $\dot{p} = f(p)$, and consider that the total population is $N = K$ at $t_H^-$ (if $t_H > 0$) and $T^-$ (if $t_H < T$).

These formal derivations should be proved carefully (using precise estimated on rescaled quantities of the form $\mathbf{n}^\epsilon(t_* + \epsilon t)$), but already suggest that the investigation of qualitative properties of minimizers (such as Proposition 10.3 and Conjecture 13.3), using indirect method, can result in dramatic simplification of the optimization problem, reducing effectively to finite dimension in this case.

## 13.4 Stationary distributions for sexual reproduction kernels

Motivated by the problem of insecticide resistance in a mosquito population, hence with sexual reproduction, Pierre-Alexandre Bliman has suggested two modeling approaches to take into account this particular feature, seldom studied (apart from single locus genetics, see in particular [57]) in the mathematical literature although considerable theories have been developed in population genetics (dating back to the works of Fisher [86] in the 1920s and Kimura [133] in the 1960s, among others). We note however that the so-called infinitesimal model with gaussian progeny distribution and trait-independent fecundity has been developed (in particular, see the discussion of [37]).

On the first hand, we can see the trait of insecticide resistance as a continuous phenotype, by which we structure the population. This is the approach of Chapter 12. On the other hand, we can rather emphasize the genetic basis of this mechanism and structure the population by its genotype. This approach is not pursued here.

In both cases, even under then natural simplifying assumptions (constant sex-ratio and constant parameters over time) it appears that the questions of existence, stability and uniqueness of stationary distributions is still open. More precisely, the question is to derive the nature of the stationary distributions under sensible sets of assumptions on the reproduction kernels.

**Setting.** Summarizing the two situations (a population structured either in phenotype or genotype), the stationary problem reads

$$\forall x \in X, \quad \frac{1}{\rho} \iint_{X^2} K(x,y,z)n(y)n(z)dydz = R(x,\rho)n(x), \tag{13.11}$$

where $\rho = \int_X n(x)dx$. Upon defining $q(x) := n(x)/\rho$ (which is a probability measure on $X$), (13.11) reads, for some $\rho > 0$,

$$Q_\rho(q) := B_\rho(q,q) = q, \quad q \in \mathcal{M}_+^1(X), \tag{13.12}$$

where $B_\rho$ is the bilinear application $\mathcal{M}_+(X) \times \mathcal{M}_+(X) \to \mathcal{M}_+(X)$ defined by

$$B_\rho(q,q)(x) = \iint_{X^2} \frac{K(x,y,z)}{R(x,\rho)} q(y)q(z)dydz.$$

Let us denote by $\mathcal{E}_\rho$ the set of solutions to (13.12), and $\mathcal{E} := \cup_{\rho>0}\mathcal{E}(\rho)$.

**Question.** Under what assumptions is $\mathcal{E}$ non-empty/a singleton?

**Tentative answer.** First, we tackle the existence problem. Assume that $R(\,\cdot\,,\rho)$ is uniformly increasing, in the sense that $\partial_\rho R \geq \nu > 0$ for all $(x,\rho) \in X \times \mathbb{R}$. For $q \in \mathcal{M}_+^1(X)$, let $\rho^*(q)$ be the unique real number such that $\iiint_{X^3} \frac{K(x,y,z)}{R(x,\rho^*(q))} q(y)q(z)dydzdx = 1$. We note that $Q_\rho^* : q \mapsto \iint_{X^2} \frac{K(x,y,z)}{R(x,\rho^*(q))} q(y)q(z)dydz$ is a well-defined and continuous mapping of the closed convex set $\mathcal{M}_+^1(X)$ into itself. If its image is compact then Schauder's fixed-point theorem applies and yields the existence of a fixed point $q^*$. If $\rho^*(q^*) > 0$ then we get a solution to (13.12).

Under mild assumptions on $K$ and $R$, one can guarantee that $\rho^* > 0$. Image compactness is more difficult - but holds for instance if $X$ itself is compact and $Q_\rho^*$ is continuous for the topology of the weak convergence of measures.

Then, we consider the question of uniqueness. Let $\mathcal{N}_1(\rho) := \{q, \int Q_\rho(q) = 1\}$. A solution $q$ to (13.12) must belong to $\mathcal{N}_1(\rho)$ and therefore also to $\mathcal{N}_2(\rho) := \mathcal{N}_1(\rho) \cap Q_\rho^{-1}(\mathcal{N}_1(\rho))$. Iterating this argument one can define the decreasing sequence $\big(\mathcal{N}_k(\rho)\big)_k$ of subsets of $\mathcal{M}_+^1(X)$ by the relationship $\mathcal{N}_{k+1}(\rho) := \mathcal{N}_k(\rho) \cap Q_\rho^{-1}(\mathcal{N}_k(\rho))$, so that

$$\mathcal{E}_\rho \subseteq \bigcap_{k\geq 1} \mathcal{N}_k(\rho) =: \mathcal{N}_\infty(\rho).$$

We can show that $\mathcal{N}_k(\rho)$ is empty for any $\rho \notin (\underline{\rho}_k, \overline{\rho}_k)$, where $(\underline{\rho}_k)_k$ (resp. $(\overline{\rho}_k)_k$) is an increasing (resp. decreasing) sequence, without the need for further assumptions. Let

$$\underline{\rho}_\infty = \lim_{k\to+\infty} \underline{\rho}_k \leq \overline{\rho}_\infty = \lim_{k\to+\infty} \overline{\rho}_k.$$

Under what assumptions can we get $\underline{\rho}_\infty = \overline{\rho}_\infty =: \rho^*$? And that $\mathcal{N}_\infty(\rho^*)$ is a singleton?

We end up with an additional remark: let $\widetilde{K}_\rho(y,z) := \int_X \frac{K(x,y,z)}{R(x,\rho)}dx$. Assume that

$$\forall \rho > 0, \quad \forall \xi \in \mathcal{M}(X), \quad \iint \widetilde{K}_\rho(y,z)\xi(y)\xi(z)dydz > 0.$$

Then $q \mapsto \int_X Q_\rho(q)(x)dx$ is strictly concave in $\mathcal{M}_+^1(X)$, therefore it reaches its minimum at some Dirac mass $q = \delta_y$ (*i.e.*, at an extreme point in the convex set $\mathcal{M}_+^1(X)$), where it is equal to $\int_X \frac{K(x,y,y)}{R(x,\rho)}dx$, and has a unique maximum point, if it reaches its maximum at $q^M$ in the interior of $\mathcal{M}_+^1(X)$. In this case, the first-order optimality conditions imply that

$$B^M : z \mapsto \iint_{X^2} \frac{K(x,y,z) + K(x,z,y)}{R(x,\rho)} q^M(y)dxdy$$

is constant, equal to some $B_*^M$, on $supp(q^M)$, and that is satisfies $B^M(z) \leq B_*^M$ for all $z \in X \backslash supp(q^M)$. One interest of this remark is that if $q$ is a steady state then $\int_X Q_\rho(q)(x)dx = 1$.

# Conclusion

## Overview

The works gathered in this thesis were motivated by the issue of vector mosquito populations control. Much attention was devoted to two innovative techniques: population replacement strategies (using cytoplasmic incompatibility and pathogen interference inducing *Wolbachia* strains) and incompatible or sterile insect techniques (SIT/IIT) for population reduction or elimination (using cytoplasmic incompatibility inducing *Wolbachia* strains and/or irradiation).

These pest management techniques share a main advantage: they are species-specific, and therefore have in principle limited impact on the environment (especially compared with the use of chemical control), although the ecological consequences in case of eradication following the use of SIT/IIT must be taken into account. Moreover, both of them rely on field releases of lab- or factory-reared individuals.

Compelling pilot or field trials have proved the feasibility and efficiency of these techniques in specific contexts, in particular targeting urban *Aedes aegypti* (primary vector of dengue) with population replacement and (small) island population of *Aedes polynesiensis* with IIT. This fact, and the outstanding importance of various species in *Aedes* genus for human health (as vectors of arboviral diseases such as dengue) have justified the focusing of our modeling effort towards this genus (presented in Chapter 3).

Our modeling viewpoint (presented in Chapter 4) was determined by the fact that acquiring detailed and reliable data on mosquito populations is a hard task. Therefore the models we developed and studied were not judged on their ability to fit necessarily incomplete data but rather on their explanatory power regarding first the mechanisms involved in qualitative observations and secondly the relative parameter importance in determining qualitative properties or quantitative values of practical interest, such as release protocol design.

In Part II we have built upon an existing scalar reaction-diffusion model used to get an intuition of *Wolbachia* dispersal in population replacement. This model represents only the frequency of *Wolbachia* infection in population in time and space. We have shown (Chapter 5) that it can be derived from a more realistic two-populations system. Then, we have analyzed two issues:

- how to ensure the success (*i.e.* the local establishment of infection) of a release protocol? (Chapter 7)
- how can the infection propagation be stopped by environmental hindrances, and can this blocking be alleviated? (Chapter 6)

In Part III, we have studied several nonlinear aspects of time dynamics for mosquito populations, neglecting the spatial dimensions. In an exploratory work, we established simple conditions for population sustainability in general seasonal systems (Chapter 11). In the remainder of the models, parameters were assumed to be constant over time. We proved (Chapter 8) that egg hatching enhancement by larvae can sustain synchronization and thus population size oscillations, which opens a possible explanation to unexpected variations (*i.e.* which do not seem to be related with environmental conditions) in abundance monitoring from trapping data. Using a mean-field model we established new criteria for population elimination using SIT/IIT (Chapter 9) and estimations on the protocol duration. This was made possible only by introducing an Allee effect in the model, which changes completely the nature of the problem compared with previous works on this topic. We introduced an optimal control viewpoint for population replacement (Chapter 10) with sensible criteria and constraints, suited to the practical application rather than to the available mathematical tools. We established that a fairly simple optimal control problem on the *Wolbachia*

infection frequency can be seen as a natural limit for the general protocol optimization problem. We also explained how the coarse recommendation that comes from the reduced problem (release as many mosquitoes as possible as soon as possible) must be altered in general.

In Part IV, we built a framework and proved preliminary results for a selection model with sexual reproduction (Chapter 12). A mosquito population does not only vary in time and space of course, but also in phenotype structure. The question leading this particular research is to understand the impact of control measures (such as mosquito releases or use of insecticide) on a key phenotype: insecticide resistance.

## Directions

Mathematically, a series of open questions have emerged during the thesis, several of which are gathered in Chapter 13. Any one of these selected problems could help understanding issues relevant for the applications, be it the design of a release protocol area, the support of a struggling *Wolbachia* infection wave or the optimization of the releases (in time) for population replacement strategies, or a finer description of insecticide resistance patterns.

More importantly, key practical questions raised by population replacement and SIT/IIT cannot be answered in a simple and unequivocal way by mathematical models alone. The quest for tighter connections between mathematicians (especially in the biomathematics community) and entomologists must be carried on. Two clear directions appear:

- the mathematical community needs challenging feedback from entomologists in order to identify and construct together mathematical problems in line with the current concerns in vector control,
- vector control programs can benefit greatly from well-understood mathematical models, for instance to help identifying experiments and measurements that are the most relevant for protocol monitoring, or optimization.

At the end of this work, let us wish that this contribution may help bridging gaps between scientific communities which most certainly have a lot to gain from mutualism.

# Bibliography


[1] *Maple 18.*, Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario.

[2] *Etymologia: Aedes aegypti*, Emerging Infectious Disease journal, 22 (2016), p. 1807.

[3] http://www.cdc.gov/zika/transmission/index.html, 2016.

[4] http://mosquito-taxonomic-inventory.info/, 2018.

[5] H. N. A., W. da Costa Silva, P. J. Leite, J. M. Gonçalves, L. P. Lounibos, and R. Lourenço-de Oliveira, *Dispersal of Aedes aegypti and Aedes albopictus (Diptera: Culicidae) in an Urban Endemic Dengue Area in the State of Rio de Janeiro, Brazil*, Mem Inst Oswaldo Cruz, 98(2) (2003), pp. 191–198.

[6] P. Adkisson and J. Tumlinson, *EDWARD F. KNIPLING 1909–2000 A Biographical Memoir*, Biographical Memoirs (National Academy of Sciences), 83 (2003).

[7] L. Alphey, *Genetic control of mosquitoes*, Annual Review of Entomology, 59 (2014), pp. 205–224.

[8] L. Alphey, A. McKemey, D. Nimmo, O. M. Neira, R. Lacroix, K. Matzen, and C. Beech, *Genetic control of Aedes mosquitoes*, Pathogens and Global Health, 107 (2013), pp. 170–179.

[9] R. Anguelov, Y. Dumont, and J. M. Lubuma, *Mathematical modeling of sterile insect technology for control of anopheles mosquito*, Comput. Math. Appl., 64 (2012), pp. 374–389.

[10] ——, *On nonstandard finite difference schemes in biosciences*, AIP Conf. Proc., 1487 (2012), pp. 212–223.

[11] R. Anguelov and J. M. Lubuma, *Contributions to the mathematics of the nonstandard finite difference method and applications*, Numerical Methods for Partial Differential Equations, 17 (2001), pp. 518–543.

[12] T. H. Ant, C. S. Herd, V. Geoghegan, A. A. Hoffmann, and S. P. Sinkins, *The Wolbachia strain wAu provides highly efficient virus transmission blocking in Aedes aegypti*, PLoS pathogens, 14 (2018), p. e1006815.

[13] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Advances in Mathematics, 30 (1978), pp. 33 – 76.

[14] G. Aronsson and I. Mellander, *A deterministic model in biomathematics. Asymptotic behavior and threshold conditions*, Mathematical Biosciences, 49.

[15] C. Atyame, P. Labbé, F. Rousset, M. Beji, P. Makoundou, O. Duron, E. Dumas, N. Pasteur, A. Bouattour, P. Fort, and M. Weill, *Stable coexistence of incompatible Wolbachia along a narrow contact zone in mosquito field populations*, Mol Ecol, 24(2) (2015), pp. 508–521.

[16] V. R. Aznar, M. S. De Majo, S. Fischer, D. Francisco, M. A. Natiello, and H. G. Solari, *A model for the development of Aedes (Stegomyia) aegypti as a function of the available food*, Journal of Theoretical Biology, 365 (2015), pp. 311 – 324.

[17] V. R. Aznar, M. Otero, M. S. De Majo, S. Fischer, and H. G. Solari, *Modeling the complex hatching and development of Aedes aegypti in temperate climates*, Ecological Modelling, 253 (2013), pp. 44 – 55.

[18] N. Bacaër, *Histoires de mathématiques et de populations*, Éditions Cassini, Paris, 2009.

[19] ——, *Sur le modèle stochastique SIS pour une épidémie dans un environnement périodique*, Journal of Mathematical Biology, 71 (2015), pp. 491–511. (french).

[20] N. Bacaër and N. Ait Dads, *Sur l'interprétation biologique d'une définition du paramètre $R_0$ pour les modèles périodiques de populations*, Journal of Mathematical Biology, 65 (2012), pp. 601–621. (french).

[21] D. Bainov and P. Simeonov, *Impulsive Differential Equations: Periodic Solutions and Applications*, vol. 66, CRC Press, 1993.

[22] F. Baldacchino, B. Caputo, F. Chandre, A. Drago, A. della Torre, F. Montarsi, and A. Rizzoli, *Control methods against invasive Aedes mosquitoes in Europe: a review*, Pest Management Science, 71 (2015), pp. 1471–1485.

[23] E. J. Balder, *On equivalence of strong and weak convergence in $L_1$-spaces under extreme point conditions*, Israel J. Math., 75 (1991), pp. 21–47.

[24] J. Bara, Z. Rapti, C. E. Cáceres, and E. J. Muturi, *Effect of larval competition on extrinsic incubation period and vectorial capacity of Aedes albopictus for dengue virus*, PLoS ONE, 10 (2015), pp. 1–18.

[25] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, vol. 17 of Mathématiques & Applications (Berlin) [Mathematics & Applications], Springer-Verlag, Paris, 1994.

[26] N. Barton, *The effects of linkage and density-dependent regulation on gene flow*, Heredity, 57 (1986), pp. 415–426.

[27] N. H. Barton and G. Hewitt, *Adaptation, speciation and hybrid zones*, Nature, 341 (1989), pp. 497–503.

[28] N. H. Barton and S. Rouhani, *The probability of fixation of a new karyotype in a continuous population*, Evolution, 45(3) (1991), pp. 499–517.

[29] N. H. Barton and M. Turelli, *Spatial Waves of Advance with Bistable Dynamics: Cytoplasmic and Genetic Analogues of Allee Effects*, The American Naturalist, 178 (2011), pp. E48–E75.

[30] H. Berestycki, B. Nicolaenko, and B. Scheurer, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16(6) (1985), pp. 1207–1242.

[31] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Society for Industrial and Applied Mathematics, 1994.

[32] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, *The global distribution and burden of dengue*, Nature, 496 (2013), pp. 504–507.

[33] S. Billiard, P. Collet, R. Ferrière, S. Méléard, and V. C. Tran, *Stochastic dynamics for adaptation and evolution of microorganisms*, ArXiv e-prints, (2016).

[34] M. S. C. Blagrove, C. Arias-Goeta, C. Di Genua, A.-B. Failloux, and S. P. Sinkins, *A Wolbachia wMel transinfection in Aedes albopictus is not detrimental to host fitness and inhibits chikungunya virus*, PLoS Neglected Tropical Diseases, 7 (2013), p. e2152.

[35] P.-A. Bliman, M. S. Aronna, F. C. Coelho, and M. A. H. B. da Silva, *Ensuring successful introduction of Wolbachia in natural populations of Aedes aegypti by means of feedback control*, Journal of Mathematical Biology, 76 (2018), pp. 1269–1300.

[36] P. A. Bliman and N. Vauchelet, *Establishing traveling wave in bistable reaction-diffusion system by feedback*, IEEE Control Systems Letters, 1 (2017), pp. 62–67.

[37] T. Bourgeron, V. Calvez, J. Garnier, and T. Lepoutre, *Existence of recombination-selection equilibria for sexual populations*, ArXiv e-prints, (2017).

[38] K. Bourtzis, *Wolbachia- based technologies for insect pest population control*, 627 (2008), pp. 104–13.

[39] A. Braides, *A handbook of* $\Gamma$*-convergence*, in Handbook of Differential Equations: Stationary Partial Differential Equations, M. Chipot and P. Quittner, eds., vol. 3, North-Holland, 2006, pp. 101 – 213.

[40] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Mathématiques appliquées pour le master, Editions Dunod, 1999.

[41] M. G. Bulmer, *The mathematical theory of quantitative genetics*, The Clarendon Press, Oxford University Press, New York, 1980. Oxford Science Publications.

[42] R. Bürger, *The mathematical theory of selection, recombination, and mutation*, Wiley Series in Mathematical and Computational Biology, John Wiley & Sons, Ltd., Chichester, 2000.

[43] F. Campillo, N. Champagnat, and C. Fritsch, *On the variations of the principal eigenvalue with respect to a parameter in growth-fragmentation models*, Communications in Mathematical Sciences, 15(7).

[44] V. O. Campo-Duarte D. E., Cardona-Salgado D., *Establishing wMelPop Wolbachia infection among wild Aedes aegypti females by optimal control approach*, Appl. Math. Inf. Sci. 1, ((2017)), pp. 1–17.

[45] C. Carrère, *Optimization of an in vitro chemotherapy to avoid resistant tumours*, J. Theoret. Biol., 413 (2017), pp. 24–33.

[46] T. J. Case and M. L. Taper, *Interspecific Competition, Environmental Gradients, Gene Flow, and the Coevolution of Species' Borders*, The American Naturalist, 155(5) (2000), pp. 583–605.

[47] E. Caspari and G. Watson, *On the evolutionary importance of cytoplasmic sterility in mosquitoes*, Evolution, 13 (1959), pp. 568–570.

[48] D. D. Chadee, P. S. Corbet, and J. J. D. Greenwood, *Egg-laying yellow fever mosquitoes avoid sites containing eggs laid by themselves or by conspecifics*, Entomologia Experimentalis et Applicata, 57 (1990), pp. 295–298.

[49] E. Chambers, L. Hapairai, B. A. Peel, H. Bossin, and S. Dobson, *Male Mating Competitiveness of a Wolbachia-Introgressed Aedes polynesiensis Strain under Semi-Field Conditions*, 5 (2011), p. e1271.

[50] M. H. T. Chan and P. S. Kim, *Modeling a Wolbachia Invasion Using a Slow–Fast Dispersal Reaction–Diffusion Approach*, Bull Math Biol, 75 (2013), pp. 1501–1523.

[51] G. Chapuisat and R. Joly, *Asymptotic profiles for a traveling front solution of a biological equation*, Math. Mod. Methods Appl. Sci., 21(10) (2011), pp. 2155–2177.

[52] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[53] J.-L. Chern, Y.-L. Tang, C.-S. Lin, and J. Shi, *Existence, uniqueness and stability of positive solutions to sublinear elliptic systems*, Proc. Roy. Soc. Edinburgh Sect. A, 141 (2011), pp. 45–64.

[54] M. Cheung, *Pairwise comparison dynamics for games with continuous strategy space*, J. Econ. Theory, 153 (2014), pp. 344–375.

[55] ——, *Imitative dynamics for games with continuous strategy space*, Games and Economic Behavior, 99 (2016), pp. 206–223.

[56] J. Clairambault, S. Gaubert, and B. Perthame, *An inequality for the Perron and Floquet eigenvalues of monotone differential systems and age structured equations*, Comptes Rendus Mathematique, 345(10) (2007), pp. 549–554.

[57] P. Collet, S. Méléard, and J. A. J. Metz, *A rigorous model study of the adaptive dynamics of Mendelian diploids*, J. Math. Biol., 67 (2013), pp. 569–607.

[58] C. Conley and R. Gardner, *An application of the generalized Morse index to traveling wave solutions of a competitive reaction-diffusion model*, University of Wisconsin-Madison. Mathematics research center. Technical summary report, 2144 (1980).

[59] C. Coron, M. Costa, H. Leman, and C. Smadi, *A stochastic model for speciation by mating preferences*, Journal of Mathematical Biology, 76 (2018), pp. 1421–1463.

[60] P. R. Crain, J. W. Mains, E. Suh, Y. Huang, P. H. Crowley, and S. L. Dobson, *Wolbachia infections that reduce immature insect survival: Predicted impacts on population replacement*, BMC Evolutionary Biology, 11 (2011), pp. 1–10.

[61] C. Curtis and T. Adak, *Population replacement in culex fatigans by means of cytoplasmic incompatibility: 1. laboratory experiments with non-overlapping generations*, Bulletin of the World Health Organization, 51 (1974), pp. 249–255.

[62] T. J. Davis, P. E. Kaufman, J. A. Hogsette, and D. L. Kline, *The Effects of Larval Habitat Quality on Aedes albopictus Skip Oviposition*, Journal of the American Mosquito Control Association, 31 (2015), pp. 321–328. doi: 10.2987/moco-31-04-321-328.1.

[63] L.-A. de Bougainville, *Voyage autour du monde par la frégate la Boudeuse et la flûte l'Etoile*, La Découverte, Paris, 1997 (1771 pour le première édition chez Saillant et Nyon, libraires, Paris).

[64] L. Desvillettes, P. E. Jabin, S. Mischler, and G. Raoul, *On selection dynamics for continuous structured populations*, Communications in Mathematical Sciences, 6(3) (2008), pp. 729–747.

[65] U. Dieckmann and M. Doebeli, *On the origin of species by sympatric speciation*, Nature, 400 (1999), pp. 354–357.

[66] O. Diekmann, J. Heesterbeek, and J. Metz, *On the definition and the computation of the basic reproduction ratio $R_0$ in models for infectious diseases in heterogeneous populations*, Journal of Mathematical Biology, 28 (1990), pp. 365–382.

[67] O. Diekmann, P.-E. Jabin, S. Mischler, and B. Perthame, *The dynamics of adaptation: an illuminating example and a Hamilton-Jacobi approach*, Theor. Popul. Biol., 67 (2005), pp. 257–271.

[68] M. Doebeli, H. J. Blok, O. Leimar, and U. Dieckmann, *Multimodal pattern formation in phenotype distributions of sexual populations*, Proc. R. Soc. B, 274 (2007), pp. 347–357.

[69] Y. Du, *Order structure and topological methods in nonlinear partial differential equations. Vol. 1*, vol. 2 of Series in Partial Differential Equations and Applications, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2006. Maximum principles and applications.

[70] Y. Du and H. Matano, *Convergence and sharp thresholds for propagation in nonlinear diffusion problems.*, J. Eur. Math. Soc., 12 (2010), pp. 279–312.

[71] C. Dufourd, *Spatio-temporal mathematical models of insect trapping : analysis, parameter estimation and applications to control*, PhD thesis, University of Pretoria, 2017.

[72] C. Dufourd and Y. Dumont, *Impact of environmental factors on mosquito dispersal in the prospect of sterile insect technique control*, Comput. Math. Appl., 66 (2013), pp. 1695–1715.

[73] Y. Dumont and J. M. Tchuenche, *Mathematical studies on the sterile insect technique for the Chikungunya disease and Aedes albopictus*, Journal of Mathematical Biology, 65 (2012), pp. 809–855.

[74] R. Durrett and S. A. Levin, *The importance of being discrete (and spatial)*, 46 (1994), pp. 363–394.

[75] H. L. C. Dutra, L. M. Barbosa dos Santos, E. P. Carsagata, J. B. L. Silva, D. A. M. Villela, R. Maciel-de Freitas, and L. A. Moreira, *From Lab to Field: The Influence of Urban Landscapes on the Invasive Potential of Wolbachia in Brazilian Aedes Aegypti Mosquitoes*, PLoS Negl Trop Dis, 9 (4) (2015).

[76] G. Duvallet, D. Fontenille, and V. Robert, *Entomologie médicale et vétérinaire*, Référence, IRD Editions/Quae, 2017.

[77] C. Dye, *The analysis of parasite transmission by bloodsucking insects*, 37 (1992), pp. 1–19.

[78] J. Edgerly and M. Marvier, *To hatch or not to hatch? Egg hatch response to larval density and to larval contact in a treehole mosquito*, Ecological entomology, 17 (1992), pp. 28–32.

[79] P. Erdos and A. Rényi, *On a classical problem of probability theory*, Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei, 6 (1961), pp. 215–220.

[80] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems*, Society for Industrial and Applied Mathematics, 2002.

[81] J. Z. Farkas, S. A. Gourley, R. Liu, and A.-A. Yakubu, *Modelling Wolbachia infection in a sex-structured mosquito population carrying West Nile virus*, Journal of Mathematical Biology, 75 (2017), pp. 621–647.

[82] J. Z. Farkas and P. Hinow, *Structured and Unstructured Continuous Models for Wolbachia Infections*, Bulletin of Mathematical Biology, 72 (2010), pp. 2067–2088.

[83] A. Fenton, K. N. Johnson, J. C. Brownlie, and G. D. D. Hurst, *Solving the Wolbachia paradox: modeling the tripartite interaction between host, Wolbachia, and a natural enemy*, The American Naturalist, 178 (2011), pp. 333–342.

[84] P. C. Fife, *Mathematical aspects of reacting and diffusing systems*, vol. 28 of Lecture notes in biomathematics, Springer-Verlag, 1979.

[85] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Archive for Rational Mechanics and Analysis, 65 (1977), pp. 335–361.

[86] R. A. Fisher, *Xxi.—on the dominance ratio*, Proceedings of the royal society of Edinburgh, 42 (1923), pp. 321–341.

[87] R. A. Fisher, *The wave of advance of advantageous genes*, Annals of Eugenics, 7 (1937), pp. 355–369.

[88] D. A. Focks, D. G. Haile, E. Daniels, and G. A. Mount, *Dynamic Life Table Model of a Container-Inhabiting Mosquito, Aedes aegypti (L.) (Diptera: Culicidae). Part 1. Analysis of the Literature and Model Development*, Journal of Medical Entomology, 30 (1993), pp. 1003–1017.

[89] R. Fourer, *AMPL : a modeling language for mathematical programming*, San Francisco, Calif. : Scientific Pr., San Francisco, Calif., 2. ed. ed., 1996.

[90] J.-P. Françoise, *Oscillations en biologie, Analyse qualitative et modèle*, Springer, 2005.

[91] R. A. Gardner, *Existence and stability of travelling wave solutions to competition models: a degree theoretic approach*, J. Diff. Equations, 44 (1982), pp. 343–364.

[92] J. Garnier, T. Giletti, F. Hamel, and L. Roques, *Inside dynamics of pulled and pushed fronts*, 98 (2011).

[93] J. Garnier, L. Roques, and F. Hamel, *Success rate of a biological invasion in terms of the spatial distribution of the founding population*, Bulletin of Mathematical Biology, 74 (2012), pp. 453–473.

[94] C. Garret-Jones, *Prognosis for interruption of malaria transmission through assessment of the mosquito's vectorial capacity*, Nature, 204 (1964), pp. 1173–1175.

[95] S. Gaubert and T. Lepoutre, *Discrete limit and monotonicity properties of the Floquet eigenvalue in an age structured cell division cycle model*, Journal of Mathematical Biology, 71 (6) (2015), pp. 1663–1703.

[96] L. Girardin and G. Nadin, *Travelling waves for diffusive and strongly competitive systems: Relative motility and invasion speed*, European Journal of Applied Mathematics, 26 (2015), p. 521–534.

[97] G. Guzzetta, F. Montarsi, F. A. Baldacchino, M. Metz, G. Capelli, A. Rizzoli, A. Pugliese, R. Rosà, P. Poletti, and S. Merler, *Potential risk of dengue and chikungunya outbreaks in northern Italy based on a population model of Aedes albopictus (Diptera: Culicidae)*, PLoS Neglect Trop D, 10 (2016), pp. 1–21.

[98] E. Hairer, C. Lubich, and M. Roche, *Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations*, BIT, 28 (1988), pp. 678–700.

[99] F. Hamel, *Reaction-diffusion problems in cylinders with no invariance by translation. part ii: Monotone perturbations*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 14 (1997), pp. 555 – 596.

[100] P. A. Hancock and H. C. J. Godfray, *Modelling the spread of wolbachia in spatially heterogeneous environments*, Journal of The Royal Society Interface, (2012).

[101] P. A. Hancock, S. P. Sinkins, and H. C. J. Godfray, *Population dynamic models of the spread of Wolbachia*, The American Naturalist, 177 (2011), pp. 323–333.

[102] ——, *Strategies for introducing Wolbachia to reduce transmission of mosquito-borne diseases*, PLoS Negl Trop Dis, 5 (2011), pp. 1–10.

[103] P. A. Hancock, V. L. White, A. G. Callahan, C. H. Godfray, A. A. Hoffmann, and S. A. Ritchie, *Density-dependent population dynamics in Aedes aegypti slow the spread of wMel Wolbachia*, Journal of Applied Ecology, 53 (2016), pp. 785–793.

[104] S. Hanson and G. J. Craig, *Cold acclimation, diapause, and geographic origin affect cold hardiness in eggs of Aedes albopictus (Diptera: Culicidae)*, J Med Entomol, 31(2) (1994), pp. 192–201.

[105] L. K. Hapairai, *Studies on Aedes polynesiensis introgression and ecology to facilitate lymphatic filariasis control*, PhD thesis, University of Oxford, 2013.

[106] L. K. Hapairai, J. Marie, S. P. Sinkins, and H. Bossin, *Effect of temperature and larval density on Aedes polynesiensis (Diptera: Culicidae) laboratory rearing productivity and male characteristics*, 132 (2013).

[107] L. K. Hapairai, M. A. C. Sang, S. P. Sinkins, and H. C. Bossin, *Population studies of the filarial vector Aedes polynesiensis (Diptera: Culicidae) in two island settings of French Polynesia*, Journal of medical entomology, 50 (2013), pp. 965–976.

[108] R. Harbach, *The Culicidae (Diptera): A Review Of Taxonomy, Classification And Phylogeny*, 1668 (2007), pp. 591–638.

[109] A. Henrot and M. Pierre, *Variation et optimisation de formes*, vol. 48, Springer-Verlag Berlin Heidelberg, 2005.

[110] M. Hertig and S. B. Wolbach, *Studies on rickettsia-like micro-organisms in insects*, The Journal of medical research, 44 (1924), p. 329.

[111] D. Hilhorst, M. Iida, M. Mimura, and H. Ninomiya, *Relative compactness in $L^p$ of solutions of some $2m$ components competition-diffusion systems*, Discrete and continuous dynamical systems, 21 (2008), pp. 233–244.

[112] D. Hilhorst, S. Martin, and M. Mimura, *Singular limit of a competition-diffusion system with large interspecific interaction*, J. Math. Anal. Appl., 390 (2012), pp. 2488–513.

[113] M. W. Hirsch, *The Dynamical Systems approach to differential equations*, Bulletin of the American Mathematical Society, 11(1).

[114] M. W. Hirsch and H. L. Smith, *Monotone dynamical systems*, in Handbook of differential equations: ordinary differential equations, vol. II, Elsevier B. V., Amsterdam, 2005, pp. 239–257.

[115] ——, *Monotone maps: a review*, Journal of Difference Equations and Applications, 11 (2005), pp. 379–398.

[116] J. Hofbauer and K. Sigmund, *Evolutionary game dynamics*, Bulletin of the American Mathematical Society, 40 (2003), pp. 479–519.

[117] A. A. Hoffmann, I. Iturbe-Ormaetxe, A. G. Callahan, B. L. Phillips, K. Billington, J. K. Axford, B. Montgomery, A. P. Turley, and S. L. O'Neill, *Stability of the wMel Wolbachia infection following invasion into Aedes aegypti populations*, PLoS Neglected Tropical Diseases, 8 (2014), pp. 1–9.

[118] A. A. Hoffmann, B. L. Montgomery, J. Popovici, I. Iturbe-Ormaetxe, P. H. Johnson, F. Muzzi, M. Greenfield, M. Durkan, Y. S. Leong, Y. Dong, H. Cook, J. Axford, A. G. Callahan, N. Kenny, C. Omodei, E. A. McGraw, P. A. Ryan, S. A. Ritchie, M. Turelli, and S. L. O'Neill, *Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission*, Nature, 476 (2011), pp. 454–457. 10.1038/nature10356.

[119] N. Honorio, C. Codeço, F. Alves, M. Magalhães, and R. Lourenço-de Oliveira, *Temporal distribution of Aedes aegypti in different districts of Rio De Janeiro, Brazil, measured by two types of traps*, J Med Entomo, 46 (5) (2009), pp. 1001–1014.

[120] M. Huang, X. Song, and J. Li, *Modelling and analysis of impulsive releases of sterile mosquitoes*, Journal of Biological Dynamics, 11 (2017), pp. 147–171. PMID: 27852161.

[121] H. Hughes and N. F. Britton, *Modeling the Use of Wolbachia to Control Dengue Fever Transmission*, Bull. Math. Biol., 75 (2013), pp. 796–818.

[122] P.-E. Jabin and H. Liu, *On a non-local selection–mutation model with a gradient flow structure*, 30 (2017), pp. 4220–4238.

[123] P. E. Jabin and G. Raoul, *On Selection dynamics for competitive interactions*, Journal of Mathematical Biology, 63(3) (2011), pp. 493–517.

[124] L. Jachowski Jr et al., *Filariasis in American Samoa. Y. Bionomics of the Principal Vector, Aedes polynesiensis Marks*, American journal of hygiene, 60 (1954), pp. 186–203.

[125] V. A. Jansen, M. Turelli, and H. C. J. Godfray, *Stochastic spread of Wolbachia*, Proceedings of the Royal Society of London B: Biological Sciences, 275 (2008), pp. 2769–2776.

[126] J. Jiang, *The algebraic criteria for the asymptotic behavior of cooperative systems with concave nonlinearities*, 6(3) (1993), pp. 193–208.

[127] F. M. Jiggins, *The spread of wolbachia through mosquito populations*, PLOS Biology, 15 (2017), pp. 1–6.

[128] S. Joanne, I. Vythilingam, N. Yugavathy, C. S. Leong, M. Wong, and S. AbuBakar, *Distribution and dynamics of Wolbachia infection in Malaysian Aedes albopictus*, Acta Trop., 148 (2015), pp. 38–45.

[129] K. N. Johnson, *The impact of Wolbachia on virus infection in mosquitoes*, Viruses, 7.

[130] S. Juliano, R. G.S., R. Maciel-de Freitas, M. Castro, C. Codeço, R. Lourenço-de Oliveira, and L. Lounibos, *She's a femme fatale: low-density larval development produces good disease vectors*, Memórias do Instituto Oswaldo Cruz, 109 (2014), pp. 1070–1077.

[131] Y. Kan-on, *Parameter dependence of propagation speed of travelling waves for competition-diffusion equations*, SIAM Journal on Mathematical Analysis, 26 (1995), pp. 340–363.

[132] T. H. Keitt, M. A. Lewis, and R. D. Holt, *Allee Effects, Invasion Pinning, and Species' Borders*, The American Naturalist, 157(2).

[133] M. Kimura, *On the probability of fixation of mutant genes in a population*, Genetics, 47 (1962), pp. 713–719.

[134] J. Kingman, *A convexity property of positive matrices*, Quart. J. Math., 12 (1961), pp. 283–284.

[135] M. Kirkpatrick and N. H. Barton, *Evolution of a Species' Range*, The American Naturalist, 150(1) (1997), pp. 1–23.

[136] E. Kisdi and S. A. H. Geritz, *Adaptive Dynamics in Allele Space: Evolution of Genetic Polymorphism by Small Mutations in a Heterogeneous Environment*, Evolution, 53 (1999), pp. 993–1008.

[137] J. Koiller, M. A. da Silva, M. O. Souza, C. T. Codeço, A. Iggidr, and G. Sallet, *Aedes, Wolbachia and dengue*, Project-Team MASAIE, (2014).

[138] A. Kolmogorov, I. Petrovsky, and N. Piskunov, *Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bulletin Université d'État à Moscou (Bjul. Moskowskogo Gos. Univ., Série internationale (1937), pp. 1–26.

[139] M. G. Kreĭn and M. A. Rutman, *Linear operators leaving invariant a cone in a Banach space*, Uspehi Matem. Nauk (N. S.), 3 (1948), pp. 3–95.

[140] J. Lamboley, A. Laurain, G. Nadin, and Y. Privat, *Properties of optimizers of the principal eigenvalue with indefinite weight and Robin conditions*, Calc. Var. Partial Differential Equations, 55 (2016), pp. Art. 144, 37.

[141] R. M. Lana, T. G. S. Carneiro, N. A. Honório, and C. T. Codeço, *Seasonal and nonseasonal dynamics of aedes aegypti in rio de janeiro, brazil: Fitting mathematical models to trap data*, Acta Tropica, 129 (2014), pp. 25 – 32. Human Infectious Diseases and Environmental Changes.

[142] R. M. Lana, M. Morais, T. França Melo de Lima, T. Carneiro, L. Stolerman, J. Santos, J. Carvajal, A. Eiras, and C. Codeço, *Assessment of a trap based Aedes aegypti surveillance program using mathematical modeling*, 13 (2018), p. e0190673.

[143] H. Laven, *Eradication of Culex pipiens fatigans through Cytoplasmic Incompatibility*, Nature, 216 (1967), pp. 383 EP –.

[144] R. Lees, J. Gilles, J. Hendrichs, M. Vreysen, and K. Bourtzis, *Back to the future: the sterile insect technique against mosquito disease vectors*, 10 (2015), pp. 156–162.

[145] R. Lees, B. Knols, R. Bellini, M. Benedict, A. Bheecarry, H. Bossin, D. Chadee, J. Charlwood, R. Dabiré, L. Djogbenou, A. Egyir-Yawson, R. Gato, L. Gouagna, M. Hassan, S. Khan, L. Koekemoer, G. Lemperiere, N. C Manoukis, R. Mozuraitis, and J. Gilles, *Review: Improving our knowledge of male mosquito biology in relation to genetic control programmes*, 132S (2014), pp. S2–S11.

[146] M. Legros, M. Otero, V. Romeo Aznar, H. Solari, F. Gould, and A. L. Lloyd, *Comparison of two detailed models of Aedes aegypti population dynamics*, Ecosphere, 7 (2016). e01515.

[147] M. A. Lewis, B. Li, and H. F. Weinberger, *Spreading speed and linear determinacy for two-species competition models*, Journal of Mathematical Biology, 45 (2002), pp. 219–233.

[148] T. J. Lewis and J. P. Keener, *Wave-block in excitable media due to regions of depressed excitability*, SIAM Journal on Applied Mathematics, 61 (2000), pp. 293–316.

[149] J. Li and Z. Yuan, *Modelling releases of sterile mosquitoes with different strategies*, Journal of Biological Dynamics, 9 (2015), pp. 1–14. PMID: 25377433.

[150] P.-L. Lions and P. E. Souganidis, *Fully nonlinear stochastic partial differential equations: non-smooth equations and applications*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 735–741.

[151] T. P. Livdahl, R. K. Koenekoop, and S. G. Futterweit, *The complex hatching response of Aedes eggs to larval density*, Ecological Entomology, 9 (1984), pp. 437–442.

[152] C. C. Lord, M. E. J. Woolhous, J. A. P. Heesterbeek, and P. S. Mellor, *Vectorborne diseases and the basic reproduction number: a case study of African horse sickness*, Medical and Veterinary Entomology, 10, pp. 19–28.

[153] A. Lorz, T. Lorenzi, J. Clairambault, A. Escargueil, and B. Perthame, *Modeling the effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors*, Bull. Math. Biol., 77 (2015), pp. 1–22.

[154] A. Lorz, S. Mirrahimi, and B. Perthame, *Dirac mass dynamics in multidimensional nonlocal parabolic equations*, Comm. Partial Differential Equations, 36 (2011), pp. 1071–1098.

[155] V. Losert and E. Akin, *Dynamics of games and genes: Discrete versus continuous time*, Journal of Mathematical Biology, 17 (1983), pp. 241–251.

[156] R. Ma, R. Chen, and Y. Lu, *Positive solutions for a class of sublinear elliptic systems*, Bound. Value Probl., (2014), pp. 2014:28, 15.

[157] G. MacDonald, *The Epidemiology and Control of Malaria*, Oxford University Press, London, 1957.

[158] R. Maciel-de Freitas, R. Souza-Santos, C. T. Codeço, and R. Lourenço-de Oliveira, *Influence of the spatial distribution of human hosts and large size containers on the dispersal of the mosquito Aedes aegypti within the first gonotrophic cycle*, Medical and Veterinary Entomology, 24 (2010), pp. 74–82.

[159] P. Magal, *Mutation and recombination in a model of phenotype evolution*, J. Evol. Equ., 2 (2002), pp. 21–39.

[160] P. Magal and G. Raoul, *Dynamics of a kinetic model describing protein exchanges in a cell population*, ArXiv e-prints, (2015).

[161] L. Malaguti and C. Marcelli, *Existence and multiplicity of heteroclinic solutions for a non-autonomous boundary eigenvalue problem*, Electronic Journal of Differential Equations, (2003), pp. 1–21.

[162] J. Marsden and M. McCracken, *The Hopf Bifurcation and its Applications*, vol. 19 of Applied mathematical sciences, Springer-Verlag, 1976.

[163] H. Matano and P. Poláčik, *Dynamics of nonnegative solutions of one-dimensional reaction–diffusion equations with localized initial data. Part I: A general quasiconvergence theorem and its consequences*, Communications in Partial Differential Equations, 41 (2016), pp. 785–811.

[164] E. A. McGraw and S. L. O'Neill, *Beyond insecticides: new thinking on an ancient problem*, Nature Reviews Microbiology, 11 (2013), pp. 181–193.

[165] J. M. Medlock, K. M. Hansford, F. Schaffner, V. Versteirt, G. Hendrickx, H. Zeller, and W. Van Bortel, *A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options*, Vector borne and zoonotic diseases (Larchmont, N.Y.), 12 (2012), p. 435—447.

[166] J. Meigen, *Systematische Beschreibung der bekannten europäischen zweiflügeligen Insekten*, vol. 1, 1818.

[167] J. Meiss, *Differential Dynamical Systems*, SIAM, 2007.

[168] R. E. Mickens, *Advances in the applications of nonstandard finite difference schemes*, World Scientific Publishing, Singapore, 2005.

[169] ——, *Dynamic consistency: a fundamental principle for constructing nonstandard finite difference schemes for differential equations*, Journal of Difference Equations and Applications, 11 (2005), pp. 645–653.

[170] S. Mirrahimi, B. Perthame, and P. Souganidis, *Time fluctuations in a population model of adaptive dynamics*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 32 (2015), pp. 41 – 58.

[171] S. Mirrahimi and G. Raoul, *Dynamics of sexual populations structured by a space variable and a phenotypical trait*, Theoretical Population Biology, 84, pp. 87–103.

[172] L. A. Moreira, I. Iturbe-Ormaetxe, J. A. Jeffery, G. Lu, A. T. Pyke, L. M. Hedges, B. C. Rocha, S. Hall-Mendelin, A. Day, M. Riegler, L. E. Hugo, K. N. Johnson, B. H. Kay, E. A. McGraw, A. F. van den Hurk, P. A. Ryan, and S. L. O'Neill, *A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya, and Plasmodium*, Cell, 139 (2009), pp. 1268–1278.

[173] L. Mousson, C. Dauga, T. Garrigues, F. Schaffner, M. Vazeille, and A.-B. Failloux, *Phylogeography of Aedes (Stegomyia) aegypti (L.) and Aedes (Stegomyia) albopictus (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations*, Genetical Research, 86 (2005), p. 1–11.

[174] C. B. Muratov and X. Zhong, *Threshold phenomena for symmetric-decreasing radial solutions of reaction-diffusion equations*, Discrete Contin. Dyn. Syst., 37 (2017), pp. 915–944.

[175] J. D. Murray, *Mathematical biology. I. An introduction*, Interdisciplinary applied mathematics, Springer, New York, 2002.

[176] G. Nadin, M. Strugarek, and N. Vauchelet, *Hindrances to bistable front propagation, application to Wolbachia*, Journal of Mathematical Biology, 76(6) (2018), pp. 1489–1533.

[177] T. Nagylaki, *Conditions for existence of clines*, Genetics, 80 (1975), pp. 595–615.

[178] T. H. Nguyen, H. L. Nguyen, T. Y. Nguyen, S. N. Vu, N. D. Tran, T. N. Le, Q. M. Vien, T. C. Bui, H. T. Le, S. Kutcher, T. P. Hurst, T. T. H. Duong, J. A. L. Jeffery, J. M. Darbro, B. H. Kay, I. Iturbe-Ormaetxe, J. Popovici, B. L. Montgomery, A. P. Turley, F. Zigterman, H. Cook, P. E. Cook, P. H. Johnson, P. A. Ryan, C. J. Paton, S. A. Ritchie, C. P. Simmons, S. L. O'Neill, and A. A. Hoffmann, *Field evaluation of the establishment potential of wMelPop Wolbachia in Australia and Vietnam for dengue control*, Parasites & Vectors, 8 (2015), p. 563.

[179] L. O'Connor, C. Plichart, A. C. Sang, C. L. Brelsfoard, H. C. Bossin, and S. L. Dobson, *Open Release of Male Mosquitoes Infected with a Wolbachia Biopesticide: Field Performance and Infection Containment*, PLoS Neglected Tropical Diseases, 6 (2012), pp. 1–7.

[180] C. F. Oliva, D. Damiens, and M. Q. Benedict, *Male reproductive biology of Aedes mosquitoes*, Acta Tropica, 132 (2014), pp. S12 – S19.

[181] M. Otero, N. Schweigmann, and H. G. Solari, *A stochastic spatial dynamical model for Aedes aegypti*, Bulletin of Mathematical Biology, 70 (2008), pp. 1297–325.

[182] T. Ouyang and J. Shi, *Exact multiplicity of positive solutions for a class of semilinear problem*, Journal of Differential Equations, 146 (1998), pp. 121 – 156.

[183] ——, *Exact multiplicity of positive solutions for a class of semilinear problem, II*, Journal of Differential Equations, 158 (1999), pp. 94 – 151.

[184] E. S. Paixão, M. G. Teixeira, and L. C. Rodrigues, *Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases*, BMJ Global Health, 3 (2018), p. e000530.

[185] N. Pasteur and M. Raymond, *Insecticide resistance genes in mosquitoes: their mutations, migration, and selection in field populations*, J. Hered., 87 (1996), pp. 444–449.

[186] B. Perthame, *Transport equations in biology*, Frontiers in mathematics, Birkhäuser Basel, 2007.

[187] ——, *Parabolic equations in biology*, Lecture Notes on Mathematical Modelling in the Life Sciences, Springer International Publishing, 2015.

[188] B. Perthame and G. Barles, *Dirac concentrations in Lotka-Volterra parabolic PDEs*, Indiana Univ. Math. J., 57 (2008), pp. 3275–3301.

[189] P. Polacik, *Parabolic equations: asymptotic behavior and dynamics on invariant manifolds*, in Handbook on Dynamical Systems vol. 2, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 835–883.

[190] ——, *Threshold solutions and sharp transitions for nonautonomous parabolic equations on* $\mathbb{R}^n$, Archive for Rational Mechanics and Analysis, 199(1) (2011), pp. 69–97.

[191] ——, *Spatial trajectories and convergence to traveling fronts for bistable reaction-diffusion equations*, Contributions to nonlinear elliptic equations and systems. A tribute to Djairo Guedes de Figueiredo on the occasion of his 80th Birthday. A.N. Carvalho et al. (eds), (2015), pp. 404–423.

[192] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag New York, 1984.

[193] A. Qualteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, Springer-Verlag New York, 2000.

[194] G. Raoul, *Macroscopic limit from a structured population model to the Kirkpatrick-Barton model*, ArXiv e-prints, (2017).

[195] F. Rivière, *Ecologie de Aedes (Stegomyia) polynesiensis, Marks, 1951, et transmission de la filariose de Bancroft en Polynésie*, PhD thesis, ORSTOM, 1988.

[196] S. Rouhani and N. H. Barton, *Speciation and the "Shifting Balance" in a continuous population*, Theoretical Population Biology, 31 (1987), pp. 465–492.

[197] G. Sallet and M. A. H. B. da Silva, *Monotone dynamical systems and some models of wolbachia in aedes aegypti populations*, ARIMA, 20 (2015), pp. 145–176.

[198] W. Sandholm, *Potential games with continuous player sets*, J. Econ. Theory, 97, pp. 81–103.

[199] ——, *Population Games and Evolutionary Dynamics*, MIT Press, Cambridge, 2010.

[200] H. Schechtman and M. O. Souza, *Costly Inheritance and the Persistence of Insecticide Resistance in Aedes aegypti Populations*, PLOS ONE, 10 (2015), pp. 1–22.

[201] T. Schmidt, I. Filipović, A. A. Hoffmann, and G. Rašić, *Fine-scale landscape genomics helps explain the slow spatial spread of Wolbachia through the Aedes aegypti population in Cairns, Australia*, 120 (2018).

[202] T. Schmidt, N. H. Barton, G. Rašić, A. Turley, B. Montgomery, I. Iturbe-Ormaetxe, P. Cook, P. A. Ryan, S. Ritchie, A. A. Hoffmann, S. L. O'Neill, and M. Turelli, *Local introduction and heterogeneous spatial spread of dengue-suppressing Wolbachia through an urban population of Aedes aegypti*, 15 (2017), p. e2001894.

[203] J. G. Schraiber, A. N. Kaczmarczyk, R. Kwok, M. Park, R. Silverstein, F. U. Rutaganira, T. Aggarwal, M. A. Schwemmer, C. L. Hom, R. K. Grosberg, and S. J. Schreiber, *Constraints on the use of lifespan-shortening Wolbachia to control dengue fever*, Journal of Theoretical Biology, 297 (2012), pp. 26 – 32.

[204] D. S. Shepard, U. E. A., Y. A. Halasa, and J. D. Stanaway, *The global economic burden of dengue: a systematic analysis*, The Lancet Infectious Diseases, 16 (2016), pp. 935 – 941.

[205] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Annali di Matematica Pura ed Applicata, 146 (1986), pp. 65–96.

[206] S. P. Sinkins, *Wolbachia and cytoplasmic incompatibility in mosquitoes*, Insect Biochemistry and Molecular Biology, 34 (2004), pp. 723 – 729. Molecular and population biology of mosquitoes.

[207] H. L. Smith, *Cooperative systems of differential equations with concave nonlinearities*, Nonlinear Analysis: Theory, Methods & Applications, 10 (1986), pp. 1037 – 1052.

[208] ——, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Providence, R.I.: American Mathematical Society, 1995.

[209] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional differential equations*, Journal of Differential Equations, 93 (1991), pp. 332 – 363.

[210] K. Snow, *The names of European mosquitoes: Part 7*, European Mosquito Bulletin, 9 (2001), pp. 4–8.

[211] M. Strugarek and N. Vauchelet, *Reduction to a single closed equation for 2 by 2 reaction-diffusion systems of Lotka-Volterra type*, SIAM Journal on Applied Mathematics, 76(5) (2016), pp. 2060–2080.

[212] M. Strugarek, N. Vauchelet, and J. P. Zubelli, *Quantifying the survival uncertainty of Wolbachia-infected mosquitoes in a spatial model*, Mathematical Biosciences and Engineering, 15(4) (2018), pp. 961–991.

[213] T. Suzuki and F. Sone, *Breeding habits of vector mosquitoes of filariasis and dengue fever in Western Samoa*, 29 (1978), pp. 279–286.

[214] C. Taing, *Dynamique de concentration dans des EDPs non locales issues de la biologie*, PhD thesis, Sorbonne Université, 2018.

[215] F. Theobald, *A monograph of the Culicidae or mosquitoes*, vol. 1, British Museum (Natural History), London., 1901.

[216] R. C. A. Thomé, H. M. Yang, and L. Esteva, *Optimal control of Aedes aegypti mosquitoes by the sterile insect technique and insecticide*, Math. Biosci., 223 (2010), pp. 12–23.

[217] A. N. Tikhonov, *Systems of differential equations containing small parameters in the derivatives*, Mat. Sb. (N.S.), 31(73) (1952), pp. 575–586.

[218] E. Trélat, J. Zhu, and E. Zuazua, *Allee optimal control of a system in ecology*, Preprint HAL, (2017).

[219] J. Tufto, *Quantitative genetic models for the balance between migration and stabilizing selection*, Genet. Res., 76 (2000), pp. 285–293.

[220] M. Turelli, *Cytoplasmic incompatibility in populations with overlapping generations*, Evolution, 64 (2010), pp. 232–241.

[221] M. Turelli and N. H. Barton, *Genetic and Statistical Analyses of Strong Selection on Polygenic Traits: What, Me Normal?*, Genetics, 138 (1994), pp. 913–941.

[222] M. Turelli and A. Hoffmann, *Rapid spread of an inherited incompatibility factor in California Drosophila*, Nature, 353 (1991), pp. 440–442.

[223] ——, *Cytoplasmic incompatibility in Drosophila simulans: dynamics and parameter estimates from natural populations*, Genetics, 140 (1995), p. 1319—1338.

[224] S. Vakulenko and V. Volpert, *New effects in propagation of waves for reaction–diffusion systems*, Asymptotic Analysis, 38 (2004), pp. 11–33.

[225] P. van den Driessche and J. Watmough, *A simple SIS epidemic model with a backward bifurcation*, Journal of Mathematical Biology, 40 (2000), pp. 525–540.

[226] G. G. van Doorn and U. Dieckmann, *The Long-Term Evolution of Multilocus Traits under Frequency-Dependent Disruptive Selection*, Evolution, 60 (2006), pp. 2226–2238.

[227] F. Vavre and S. Charlat, *Making (good) use of Wolbachia: what the models say*, Current Opinion in Microbiology, 15 (2012), pp. 263 – 268. Ecology and industrial microbiology/Special section: Microbial proteomics.

[228] M. Vazeille, S. Moutailler, D. Coudrier, C. Rousseaux, H. Khun, M. Huerre, J. Thiria, J.-S. Dehecq, D. Fontenille, I. Schuffenecker, P. Despres, and A.-B. Failloux, *Two Chikungunya Isolates from the Outbreak of La Reunion (Indian Ocean) Exhibit Different Patterns of Infection in the Mosquito, Aedes albopictus*, PLoS ONE, 2 (2007), pp. 1–9.

[229] D. A. M. Villela, C. T. Codeço, F. Figueiredo, G. A. Garcia, R. Maciel-de Freitas, and C. J. Struchiner, *A Bayesian hierarchical model for estimation of abundance and spatial density of Aedes aegypti*, PLoS ONE, 10(4) (2015).

[230] A. Volpert, V. Volpert, and V. Volpert, *Traveling wave solutions of parabolic systems*, vol. 140 of Translation of Mathematical Monographs, Amer. Math. Society, Providence, 1994.

[231] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.

[232] T. Walker, P. H. Johnson, L. A. Moreira, I. Iturbe-Ormaetxe, F. D. Frentiu, C. J. McMeniman, Y. S. Leong, Y. Dong, J. Axford, P. Kriesner, A. L. Lloyd, S. A. Ritchie, S. L. O'Neill, and A. A. Hoffmann, *The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations*, Nature, 476 (2011), pp. 450–453. 10.1038/nature10355.

[233] J. H. Werren, L. Baldo, and M. E. Clark, *Wolbachia: master manipulators of invertebrate biology*, Nature Review Microbiology, 6 (2008), pp. 741–751.

[234] B. Wu and R. Cui, *Existence, uniqueness and stability of positive solutions to a general sublinear elliptic systems*, Bound. Value Probl., (2013), pp. 2013:74, 14.

[235] S.-L. Wu and W.-T. Li, *Global asymptotic stability of bistable traveling fronts in reaction-diffusion systems and their applications to biological models*, Chaos, Solitons & Fractals, 40 (2009), pp. 1229 – 1239.

[236] D. Xiao, *Dynamics and bifurcations on a class of population model with seasonal constant-yield harvesting*, Discrete and Continuous Dynamical Systems Series B, 21(2) (2016), pp. 699–719.

[237] H. Yang, *Assessing the influence of quiescence eggs on the dynamics of mosquito Aedes aegypti*, Applied Mathematics, 5 (2014), pp. 2696–2711.

[238] H. M. Yang, M. L. G. Macoris, K. C. Galvani, M. T. M. Andrighetti, and D. M. V. Wanderley, *Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue*, Epidemiology and Infection, 8 (2009), pp. 1188–1202.

[239] H. L. Yeap, P. Mee, T. Walker, A. R. Weeks, S. L. O'Neill, P. Johnson, S. A. Ritchie, K. M. Richardson, C. Doig, N. M. Endersby, and A. A. Hoffmann, *Dynamics of the "Popcorn" Wolbachia Infection in Outbred Aedes aegypti Informs Prospects for Mosquito Vector Control*, Genetics, 187 (2011), pp. 583–595.

[240] H. L. Yeap, G. Rasic, N. M. Endersby-Harshman, S. F. Lee, E. Arguni, H. Le Nguyen, and A. A. Hoffmann, *Mitochondrial DNA variants help monitor the dynamics of Wolbachia invasion into host populations*, Heredity, 116 (2016), pp. 265–276. Supplementary information available for this article at http://www.nature.com/hdy/journal/v116/n3/suppinfo/hdy201597s1.html.

[241] Z. Zhang, T. Ding, W. Huang, and Z. Dong, *Qualitative Theory of Differential Equations*, no. 101 in Translations of Mathematical Monographs, American Mathematical Society, Providence, 1991.

[242] B. Zheng, M. Tang, J. Yu, and J. Qiu, *Wolbachia spreading dynamics in mosquitoes with imperfect maternal transmission*, Journal of Mathematical Biology, 76 (1) (2018), pp. 235–263.

[243] A. Zlatos, *Sharp transition between extinction and propagation of reaction*, J. Amer. Math. Soc., 19 (2006), pp. 251–263.